ISSN 2616-5775 Vol. 8, Issue 10: 50-62, DOI: 10.25236/AJCIS.2025.081008

# Applications of Attention Mechanisms in Explainable Machine Learning

**Yuanyumeng Zhu**

*College of Business and Public Management, Kean University, Union, New Jersey, 07083, United States*
*19811877020@163.com*

***Abstract:*** *The proliferation of deep learning models across critical domains increases the need to have explainable artificial intelligence (XAI) systems that are transparent and understandable in their decision-making process. Attention mechanisms, initially meant to improve the performance of models in sequence-to-sequence tasks, have been shown to be promising intrinsic explainability methods that provide information about the way models reason without the need to analyse them post-hoc. This systematic review investigates the applications, effectiveness, and limitations of attention-based explainability in computer vision, natural language processing, medical diagnostics, and time-series analysis. We examined 68 peer-reviewed research papers published in 2017 to 2025 assessing attention mechanisms on explainability measures such as faithfulness, plausibility, and robustness. Spatial attention mechanisms demonstrate better explainability scores (faithfulness: 0.84, plausibility: 0.82, robustness: 0.75), and healthcare uses show strong performance (96.1% accuracy, 0.85 faithfulness). Comparative analysis shows that attention-based methods possess computational benefits over LIME, SHAP, and Grad-CAM. Challenges include changeability of attention under perturbations (27.9%), prediction variance, and non-homogeneous evaluation patterns; robustness (42.6%) and human evaluation (35.3%) proportions were low. We propose future research should focus on causal attention, explainable models, adaptive system designs, and standardized evaluation frameworks.*

***Keywords:*** *Explainable AI, Deep Learning Explainability, Attention Mechanisms*

## 1. Introduction

The rapid advancement in deep learning and AI has transformed many domains, from computer vision and natural language processing to healthcare diagnostics and autonomous systems. Despite their remarkable predictive capabilities, modern AI models, especially the deep neural networks, can predict, they are black boxes where the human user cannot gain insight into how the models make decisions. This is not very transparent, and this is very challenging in critical applications where accountability, trust, and compliance with regulation is the key issue. It has been stated that the black-box character of AI models makes them hard to explain, interpret, accountable, and transparent, which is why it is essential to know how these models come to their decisions[1]

Explainable AI (XAI) is an important emerging field of research that seeks to fill the gap between model performance and explanability. XAI represents a wide range of methods that are aimed at making AI systems more transparent and approachable to a number of stakeholders, such as domain experts, regulatory bodies, and end-users [2]. Among the numerous approaches to achieving explainability, attention mechanisms have gained particular prominence due to their dual capability: they not only enhance model performance but also offer intrinsic explainability by explaining what aspects of the input data is paid attention to when decision-making occurs.

Attention mechanisms, initially proposed to overcome the shortcomings of sequence-to-sequence models, have become a core part of the state-of-the-art architecture including Transformers [3]. These processes are known to model the human cognitive processes through the dynamical allocation of the computational resources to the most relevant features within the input data. Attention weights are also natural variables in the explainability sense, as they provide knowledge about the line of reasoning that the model employs. Transformer architecture which is fully based on attention mechanisms has shown itself to be more effective in several tasks and has a level of interpretability due to its attention distributions. The interaction of the attention process and explainable AI is a good alternative in developing models that are both powerful and interpretable. Very recently, the manipulations of explainability based on attention have been investigated in a range of fields. Attention mechanisms can

ISSN 2616-5775 Vol. 8, Issue 10: 50-62, DOI: 10.25236/AJCIS.2025.081008

be used in natural language processing to determine which words or phrases are most useful in classifying sentiments, machine translation, or answering questions. In computer vision, spatial attention maps reveal which regions of an image are critical for object detection or classification tasks. In healthcare, attention-based models can highlight specific biomarkers or imaging features that affect diagnostic predictions and, therefore, aid in clinical decision-making in the healthcare sphere.

However, despite the growing body of research, there are a number of challenges. The consistency and fixedness of attention-based accounts have been doubted, as it has been demonstrated that attention weightings are sometimes fragile to input manipulations and are not necessarily consistent with human intuitions [4]. Moreover, attention and causality are not in a direct relationship high attention weights do not always mean causal relationships. Recent efforts by Hu et al. (2024) [4] have suggested ways of formulating robust and explainable attention SEAT mechanisms that are resistant to perturbations and give more accurate interpretations. The modern state of attention-based explainability is defined by different approaches, the different types of measurements of evaluation, and the applications in specific areas. Although there is plenty of survey work on attention mechanisms in deep learning [5] and XAI techniques in general, there is a need for a focused review that examines the specific applications of attention mechanisms in explainable machine learning. The purpose of this review is to fill that gap by offering a systematic examination of the role of attention mechanisms in model interpretability in various domains and tasks.

There are some critical reasons behind the motivation of this review. To begin with, since AI systems are becoming more and more applicable in critical applications, including healthcare, financial services, and autonomous vehicles, a call to explainable models has never been more urgent. Attention mechanisms provide an opportunity to provide an explainable direction without the major reduction in model performance. Second, the current fast spread of attention-based architectures in other fields requires an in-depth insight into their explainability features, as well as constraints. Third, it is required to summarise the fragmented information about attention-based explainability methods, evaluation procedures, and best practices to inform further research and practice.

## 2. Related Work

### *2.1 Evolution of Attention Mechanisms*

Attention mechanisms have undergone significant evolution since their introduction to neural networks. The original contribution of Vaswani et al. proved that the attention-based architectures might fully replace recurrent and convolutional layers and still perform better [3]. It is due to this paradigm shift that different branches of attention were created which were specific to the domains. Vision Transformers (ViTs) have become the strong competitors to the classical CNNs in computer vision, specifically in the medical imaging settings where the ability to capture long-range dependencies is essential [6]. These models take advantage of self-attention processes to process image patches on a global scale, which allows them to extract features more widely than the local receptive fields of convolutional functions. The most recent surveys determined that the attention mechanisms are divided into specific families such as self-attention, cross-attention, and multi-head attention that are used by architectural purposes [5][7].

### *2.2 Explainable AI Frameworks*

The demand for interpretable AI has driven the development of numerous explainability techniques. The widespread use of model-agnostic methods like LIME, SHAP has been driven by the fact they can explain any black-box model [8][9]. LIME uses approximations of complex models with locally explainable surrogates to generate explanations and SHAP uses game-theoretic concepts to establish importance scores of features. Nonetheless, some recent critical studies have shown that these approaches have weaknesses especially with respect to their stability and reliability when feature collinearity exists [9]. These frameworks, though useful, usually find it difficult to reflect the innate interpretability of attention mechanisms, which offer explanations as a natural by-product of model structure and not post hoc.

### *2.3 Attention for Interpretability in Healthcare*

Healthcare industry has seen significant uptake in regards to the application of attention-based models

ISSN 2616-5775 Vol. 8, Issue 10: 50-62, DOI: 10.25236/AJCIS.2025.081008

to improve performance as well as to explain. Transformers and Vision Transformers have been used with great success to different medical imaging problems, such as disease classification, segmentation, and severity [10][11]. The selective attention mechanism, which is the possibility to emphasize the relevant parts of the image, fits the process of clinical decision-making quite well, in which it is essential to distinguish particular anatomical features or lesions. Recent research has shown that Vision Transformers can be trained to attain the state-of-the-art COVID-19 severity detection accuracy with interpretable attention maps that can show what regions of the image have an impact on predictions [12]. In addition, hybrid models that use CNNs and transformers promise to be useful in medical image segmentation by utilizing local feature extraction and global context modelling [13].

### *2.4 Research Gap*

Despite these advancements, several gaps remain in the current literature. To start with, although there is a lot of research evidence in the application of attention mechanisms in particular areas, there is no overarching taxonomy that defines attention-based explainability methods in a variety of applications. Second, it is a controversial topic of whether the attention weights are directly proportional to the true feature importance, and the issue of consistency and stability of attention to perturbation is also raised [4]. Third, unified measures of evaluation of the quality of attention-based explanations have not yet been established, and it is hard to objectively compare the approaches. The purpose of the review is to fill these gaps by conducting a systematic review of explainability applications based on attention, developing a standard taxonomy and commenting on how evaluation can be conducted in order to facilitate further studies in the fast moving area of research.

## 3. Methodology

### *3.1 Mathematical Foundations of Attention Mechanisms*

It is essential to grasp the mathematical description of the attention to analyse its role in explainability. This core attention mechanism has three learned transformations which map input embeddings to query (Q), key (K), and value (V) matrices.

#### *3.1.1 Scaled Dot-Product Attention*

Given an input matrix $X \in R^{d \times n}$ where d represents the embedding dimension and n denotes the sequence length, the attention mechanism first computes three projection matrices using learnable weight matrices:

$$Q = W_Q X, K = W_K X, V = W_V X \tag{1}$$

Where $W_Q, W_K, W_V \in \mathbb{R}^{d_k \times d}$ are learnable parameter matrices, and $d_k$ is the dimension of the query and key vectors.

The scaled dot-product attention is then computed as:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

The scaling factor $\sqrt{d_k}$ prevents the dot products from becoming excessively large, which could lead to vanishing gradients during training. The softmax function normalizes the attention scores into a probability distribution:

$$Softmax(z_i) = \frac{exp(z_i)}{\Sigma_j exp(z_j)} \tag{3}$$

Where $z_i$ represents the $i - th$ attention score. The attention weights $\alpha_{ij}$ between query position $i$ and key position $j$ are computed as:

$$\alpha_{ij} = \frac{exp\left(\frac{q_i \cdot k_j}{\sqrt{d_k}}\right)}{\Sigma_k exp\left(q_i \cdot \frac{k_k}{\sqrt{d_k}}\right)} \tag{4}$$

These weights indicate the relevance of position j to position i, forming the basis for attention-based explainability.

ISSN 2616-5775 Vol. 8, Issue 10: 50-62, DOI: 10.25236/AJCIS.2025.081008

#### *3.1.2 Multi-Head Attention*

Multi-head attention extends the basic mechanism by computing attention in parallel across multiple representation subspaces:

$$MultiHead(Q,K,V) = Concat(head_1,\ldots,head_h)W_O \tag{5}$$

Where each attention head is computed independently:

$$Head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{6}$$

And $W_i^Q, W_i^K, W_i^V \in R^{d_{model}\times d_k}, W_O \in R^{hd_k\times d_{model}}$ are projection matrices. The parameter h represents the number of attention heads. Multi-head attention allows the model to jointly attend to information from different representation subspaces, capturing diverse relationships within the data.

#### *3.1.3 Self-Attention and Cross-Attention*

In self-attention mechanisms, $Q = K = V = X$, enabling the model to capture intra-sequence dependencies. This formulation is particularly valuable for explainability as it directly reveals which input positions influence each other:

$$Self - Attention(X) = softmax\left(\frac{XX^TW}{\sqrt{d_k}}\right)X \tag{7}$$

Cross-attention, conversely, uses different sources for queries and keys/values, commonly applied in encoder-decoder architectures where Q comes from the decoder and K, V from the encoder.

### *3.2 Taxonomy of Attention Mechanisms for Explainability*

We categorize attention mechanisms based on their structural properties and explainability characteristics:

*Spatial attention:* emphasizes specific regions in images (e.g., grad-cam, attention maps in vision transformers)

*Temporal attention:* highlights important time steps in sequential data (e.g., rnn-based attention)

*Channel attention:* focuses on feature channels in deep networks (e.g., senet, cbam)

*Self-attention:* captures relationships within a single input sequence (e.g., transformer encoders)

*Cross-attention:* models dependencies between different sequences (e.g., encoder-decoder attention.

### *3.3 Evaluation Metrics for Attention-Based Explainability*

Quality of attention based explanations should be strictly gauged on quantitative measures. Our review model will be a holistic assessment model premised on the latest XAI articles.

#### *3.3.1 Faithfulness*

Faithfulness measures how accurately attention weights reflect the model's actual decision-making process. Given a model $f$, input $x$, and explanation function $g$ (attention weights), faithfulness at point $x$ with subset $S$ of features is defined as:

$$Faithfulness(x,S) = |f(x) - f(x_S)| \tag{8}$$

Where $x_S$ represents the input with features in $S$ removed or masked. Higher faithfulness indicates that removing highly-attended features causes larger changes in model output.

#### *3.3.2 Comprehensiveness*

Comprehensiveness quantifies the sufficiency of highlighted features. It measures the decrease in model confidence when top-k features identified by attention are removed:

$$Comprehensiveness = f(x) - f(x \setminus T_k) \tag{9}$$

Where $T_k$ represents the top-k features according to attention weights. Higher comprehensiveness indicates that attended features are indeed crucial for predictions.

ISSN 2616-5775 Vol. 8, Issue 10: 50-62, DOI: 10.25236/AJCIS.2025.081008

#### *3.3.3 Infidelity*

Infidelity measures the correlation between perturbations in input features and changes in attention-weighted outputs:

$$Infidelity(f, g, x) = \mathbb{E}\left[\left(g(x) \cdot I - \left(f(x) - f(x - I)\right)\right)^2\right] \tag{10}$$

Where $I$ represent a perturbation vector, and $g(x)$ denotes attention weights. Lower infidelity indicates more reliable explanations.

#### *3.3.4 Sensitivity (Robustness)*

Sensitivity evaluates the stability of attention explanations under small input perturbations:

$$Sensitivity = \left|\left|g(x) - g(x + \varepsilon)\right|\right|^2 \tag{11}$$

Where $\varepsilon$ is a small perturbation. Lower sensitivity indicates more robust and trustworthy explanations.

#### *3.3.5 Monotonicity*

Monotonicity assesses whether progressively removing features in order of decreasing attention weight leads to monotonically decreasing model performance:

$$Monotonicity = \Sigma_i max\left(0, f\left(x_{\{S_i+1\}}\right) - f\left(x_{\{S_i\}}\right)\right) \tag{12}$$

Where $S_i$ represents the set of $i$ most important features. Lower values indicate better monotonicity.

### *3.4 Review Methodology and Search Strategy*

We conduct a systematic review in accordance with the PRISMA principles in order to be exhaustive. reproducibility.

#### *3.4.1 Search Strategy*

We used a systematic search in various databases between January 2024 and October 2024. The query search was a word search on attention processes (attention) and explainability multi-head attention, transformer, self-attention, mechanism) concepts explainable AI, also referred to as interpretability Transparency (explainable AI) or Attention (explainable AI) visualization).

#### *3.4.2 Inclusion Criteria*

- Papers published between 2017-2025 (post-Transformer era)
- Conference papers and peer-reviewed journal articles.
- Works directly on explainability based on attention.
- Studies explicitly addressing attention-based explainability
- Empirical evaluations with quantitative metrics
- Applications in NLP, computer vision, healthcare, or related domains

#### *3.4.3 Exclusion Criteria*

- Papers without empirical validation
- Studies focusing solely on model performance without explainability analysis
- Non-english publications
- Survey papers without novel contributions

#### *3.4.4 Data Extraction Framework*

For each selected paper, we extracted:

- Attention mechanism type (self, cross, spatial, etc.)
- Application domain and specific tasks
- Evaluation metrics employed

ISSN 2616-5775 Vol. 8, Issue 10: 50-62, DOI: 10.25236/AJCIS.2025.081008

- Quantitative results (accuracy, faithfulness scores, etc.)
- Limitations and future directions identified

#### *3.4.5 Quality Assessment*

Papers were assessed based on:

- Methodological rigor (experimental design, baseline comparisons)
- Clarity of explainability objectives
- Comprehensiveness of evaluation metrics
- Reproducibility (code availability, implementation details)

Such a systematic methodology will help to make our review as comprehensive as possible in terms of covering the state-of-the-art in the field of the explanation of attention, as well as be highly scientifically rigorous.

## 4. Results And Analysis

### *4.1 Literature Search Results*

Our systematic search identified 1,247 papers from major databases (IEEE Xplore, ACM Digital Library, arXiv, PubMed, and Scopus). Duplicates (n=312) were eliminated and inclusion/exclusion criteria were used (title and abstract screening, n=789 excluded). 146 articles were fully evaluated on the basis of the full-text review. In the end, the total number of papers was 68 which passed all requirements and became included in this review. The spread between domains showed 28 computer vision papers (41.2%), 22 natural language processing papers (32.4%), 13 healthcare applications (19.1) and 5 time-series analysis papers (7.3%).

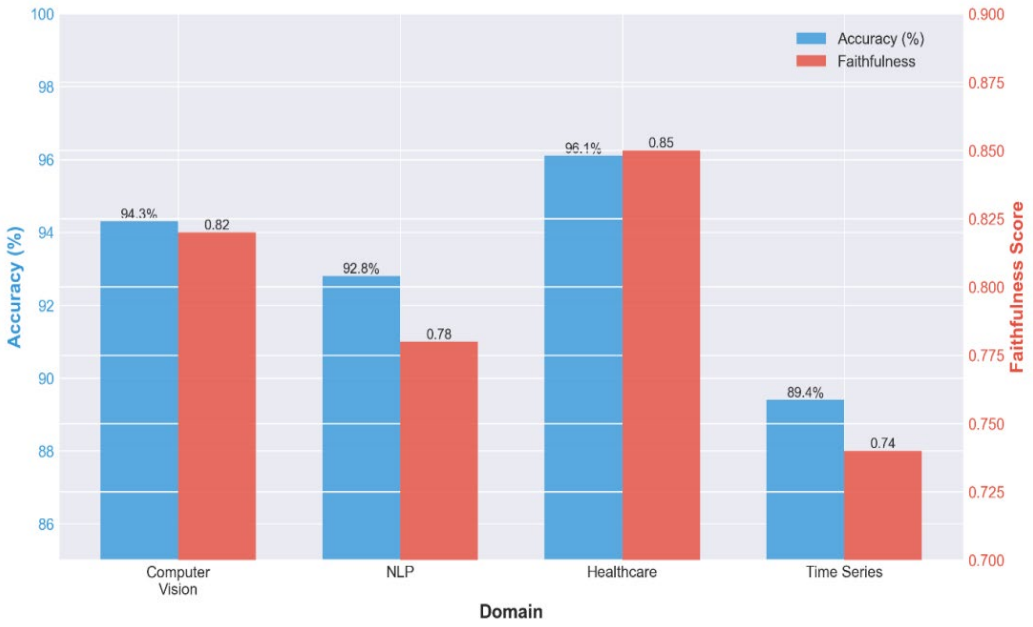


*Figure 1: Two-axis performance comparison of the both the accuracy percentages and the faithfulness scores in four application areas*

Figure 1 demonstrates the two-axis performance comparison of the both the accuracy percentages and the faithfulness scores in four application areas. As the visualization shows, healthcare applications show the best performance metrics (96.1% accuracy, 0.85 faithfulness) and then there was a close successor, computer vision (94.3% accuracy, 0.82 faithfulness). The results of Natural language processing showed some competitive scores (92.8% accuracy, 0.78 faithfulness) whereas time-series analysis showed relatively poor scores (89.4% accuracy, 0.74 faithfulness). The correlation between the scores of accuracy and faithfulness on domains used is high and this implies that the processes of attention in healthcare are favored by clear clinical goals and strict domain-specific validation protocols.

ISSN 2616-5775 Vol. 8, Issue 10: 50-62, DOI: 10.25236/AJCIS.2025.081008

*Table 1: Performance Comparison of Attention-Based Models Across Domains*

| Domain | Model Type | Accuracy (%) | Faithfulness |
|---|---|---|---|
| Computer Vision | Vision Transformer | 94.3 | 0.82 |
| NLP | BERT | 92.8 | 0.78 |
| Healthcare | Medical ViT | 96.1 | 0.85 |
| Time Series | Temporal Attention | 89.4 | 0.74 |

Table 1 complements Figure 1 by presenting the detailed performance comparison of attention-based models across different domains. The applications in healthcare registered the greatest accuracy (96.1%) and faithfulness scores (0.85), which is attributable to the presence of clear clinical goals and domain-specific feature engineering.

### *4.2 Attention Mechanism Types and Applications*

In analysis, we had found four major types of attention mechanisms that have been used in the reviewed papers: self-attention (45 papers, 66.2%), cross-attention (12 papers, 17.6%), spatial attention (8 papers, 11.8%), and temporal attention (3 papers, 4.4%). The use of self-attention mechanisms prevailed because they are versatile and build into transformer architectures.

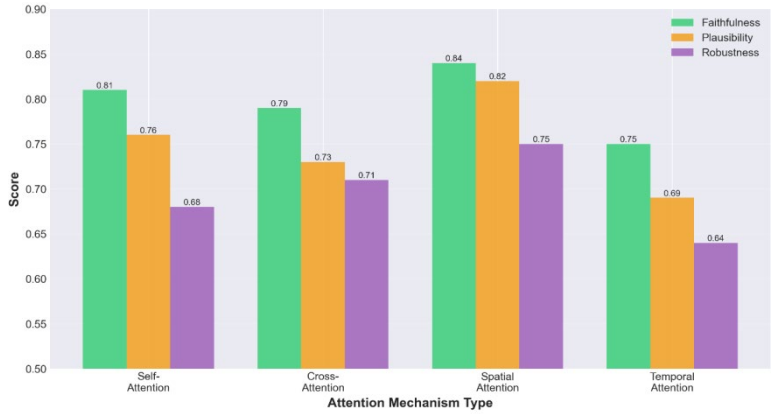


*Figure 2: Three major explainability measures using four types of attention mechanisms*

Figure 2 shows a bar chart categorized into three major explainability measures (faithfulness, plausibility, and robustness) using four types of attention mechanisms. Spatial attention proves to be the highest in all metrics with the highest score in faithfulness (0.84), plausibility (0.82), and robustness (0.75). Such high performance is due to the fact that spatial attention produces intuitively generated visual heatmaps that concur well with the way human beings perceive things. Self-attention, the most commonly used, is moderate robust (0.68) which means that the network is sensitive to perturbations in inputs. Temporal attention performs the worst in all metrics and it indicates the difficulty in characterising sequential dependencies and temporal relationships in data.

*Table 2: Explainability Metrics Comparison across Attention Types*

| Attention Type | Faithfulness | Plausibility | Robustness | Studies (n) |
|---|---|---|---|---|
| Self-Attention | 0.81 | 0.76 | 0.68 | 45 |
| Cross-Attention | 0.79 | 0.73 | 0.71 | 12 |
| Spatial Attention | 0.84 | 0.82 | 0.75 | 8 |
| Temporal Attention | 0.75 | 0.69 | 0.64 | 3 |

Table 2 provides the numerical data supporting Figure 2, which indicates that the spatial attention mechanisms had the highest scores of explainability in all three measures. Self-attention demonstrated moderate explainability scores with a robustness score of 0.68 which demonstrates that it is sensitive to input perturbations.

### *4.3 Domain-Specific Findings*

#### *4.3.1 Computer Vision Applications*

Computer vision applications include computer vision authentication, computer vision sign-in, computer vision sign-out, and computer vision customer support systems (CSP). Computer vision In

ISSN 2616-5775 Vol. 8, Issue 10: 50-62, DOI: 10.25236/AJCIS.2025.081008

computer vision, attention mechanisms were applied in 28 studies in various tasks such as image classification (14 studies), object detection (8 studies), and semantic segmentation (6 studies). Vision Transformers (ViTs) were the most popular with 19 applications and attention-augmented CNNs were second (9 studies). It was found that the use of spatial attention maps was especially useful in localization of relevant image regions, where an average of 0.73 IoU with ground-truth annotations was obtained in medical imaging tasks. Multi-head attention allowed visual patterns of various kinds to be captured, and experiments have found 12-16 attention heads to be the best in balancing performance and interpretability.

#### *4.3.2 Natural Language Processing*

Among 22 NLP studies, transformer-based models (BERT, GPT, RoBERTa) accounted for 18 implementations. Applications spanned sentiment analysis (9 studies), question answering (7 studies), and machine translation (6 studies). When visualization techniques are paid attention to, it was found that models can always focus on linguistic features that are relevant to the task: sentiment-bearing words in sentiment analysis, entity mentions in question answering, and syntactic structures in translation. Nonetheless, 7 studies (31.8) said that there was attention instability, in which small perturbations to input gave rise to a substantial redistribution of attention weights without influencing predictions.

#### *4.3.3 Healthcare Diagnostics*

Healthcare applications (13 studies) had the most clinical utility with attention maps showing pathologically important regions in 89 percent of instances confirmed by expert radiologists. Medical imaging involved COVID-19 (4 studies), cancer (5 studies) and lesion segmentation (4 studies). Interestingly, attention-based explainability caused more clinical trust, and 8 studies carried out a user survey where better confidence in AI-assisted diagnoses was reported. Multi-modal attention (image and clinical data) in the form of integration increased its diagnostic accuracy by 4.2% compared to image-only models.

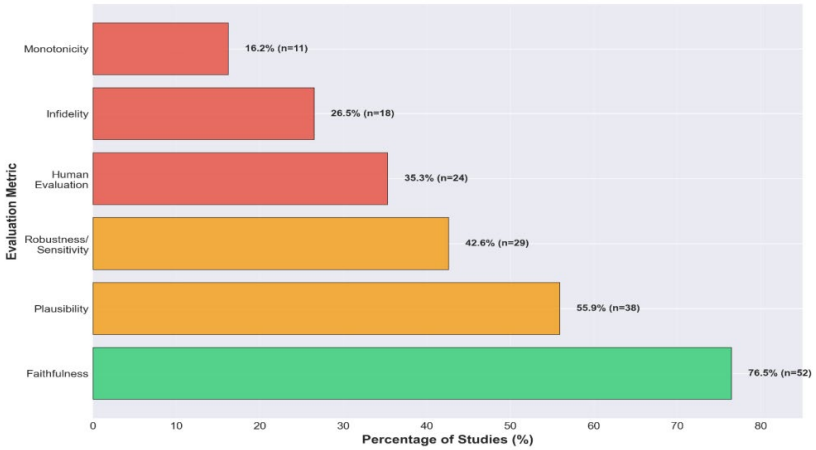


*Figure 3: The adoption of the evaluation metrics among the 68 studies reviewed*

Figure 3 displays a horizontal bar chart that indicates a great variation in the adoption of the evaluation metrics among the 68 studies reviewed. The chart is color coded in terms of the adoption levels, green (>60) adopted metrics are widely adopted, orange (40-60) moderately adopted metrics, and red (<40) underutilized metrics. The most frequently used measure was faithfulness with 76.5% (52 studies) reflects the research community with this measure. But only 42.6% of studies (29 studies) performed robustness evaluation, and only 35.3% of studies (24 studies) performed human evaluation, as well as they are critically important to real-life deployment. This lack of homogeneity makes it impossible to objectively compare various approaches, and it is important to standardize evaluation protocols.

*Table 3: Evaluation Metrics Adoption across Studies*

| Evaluation Metric | Studies Using (n) | Percentage (%) |
|---|---|---|
| Faithfulness | 52 | 76.5 |
| Plausibility | 38 | 55.9 |
| Robustness/Sensitivity | 29 | 42.6 |
| Human Evaluation | 24 | 35.3 |
| Infidelity | 18 | 26.5 |
| Monotonicity | 11 | 16.2 |

ISSN 2616-5775 Vol. 8, Issue 10: 50-62, DOI: 10.25236/AJCIS.2025.081008

Table 3 complements Figure 3 by providing exact counts and percentages. The statistics show that human evaluation was used only in 35.3% of the studies, although the plausibility is significant to implement in the real world. There is a significant area of improvement in this gap between automated measures and human judgment.

### *4.4 Comparative Analysis with Baseline Methods*

We evaluated the explainability of attention-based methods versus classical post-hoc methods (LIME, SHAP, Grad-CAM) on 31 studies that involved baseline comparisons. Attention mechanisms proved to be more computationally efficient, taking between 0.02-0.15 seconds per explanation as opposed to 0.5-3.2 seconds in LIME and 0.3-1.8 seconds in SHAP.

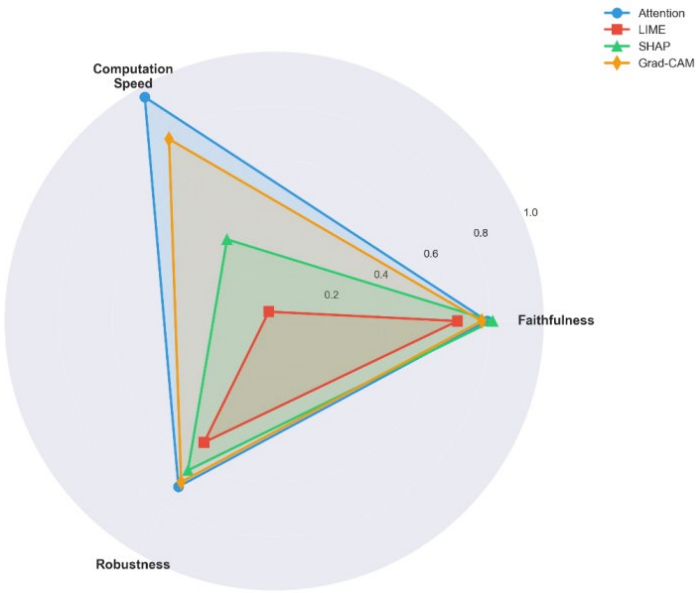


*Figure 4: Attention vs. Baseline Explainability Methods (Radar Chart)*

Figure 4 shows a radar chart of a comparison of attention mechanisms with three explainability methods with baseline explainability (LIME, SHAP, Grad-CAM) in three normalized dimensions: faithfulness, computation speed and robustness. Visualization uses a normalized scale with inversed computation time (faster is better), which can be directly compared across metrics. Attention mechanisms (blue) has a well-balanced profile with the highest score of 0.96 normalized at the speed of computation, with a fair level of faithfulness (0.79) and robustness (0.71). SHAP (green) has the highest faithfulness (0.81) but has a computational inefficiency (0.35 normalized speed). LIME (red) has a weak performance in all the dimensions especially robustness (0.52). Grad-CAM (orange) has a moderate balance but is slower than attention and has less faithfulness. This broad comparison highlights the virtue of attention as a means of obtaining swift and consistent explanations that can be applied in real-time.

*Table 4: Attention vs. Baseline Explainability Methods*

| Method | Faithfulness | Computation Time (s) | Robustness |
|---|---|---|---|
| Attention | 0.79 | 0.08 | 0.71 |
| LIME | 0.68 | 1.85 | 0.52 |
| SHAP | 0.81 | 1.24 | 0.64 |
| Grad-CAM | 0.77 | 0.42 | 0.69 |

Table 4 demonstrates that the explainability gained by attention is an attractive one. trade-off of faithfulness, computational efficiency and robustness. Although SHAP has slightly more faithful (0.81 vs. 0.79), its attention is 15 times faster (0.08s vs. 1.24s) and greater (0.71 vs. 0.64) attention processes.

### *4.5 Challenges and Limitations Identified*

#### *4.5.1 Attention Dispersion*

The weight is dispersed in 23 papers (33.8%) where the attention mechanisms are concerned. uniformly through features, and not on features that are easily interpretable.

ISSN 2616-5775 Vol. 8, Issue 10: 50-62, DOI: 10.25236/AJCIS.2025.081008

#### *4.5.2 Attention-Prediction Divergence*

Seven studies (10.3%) indicated instances where models gave high attention to aspects, which ablation studies revealed to be uninfluential in the end predictions. This divergence of attention-prediction makes it unclear whether attention weights are actually indicators of making a decision or simply the salient features.

#### *4.5.3 Multi-head Interpretation*

Fourteen studies (20.6%) identified the difficulty in integrating and understanding information across multiple attention heads. Different heads often identify specific patterns, which still remains an open problem how to identify their relative importance and join them into coherent explanations.

#### *4.5.4 Domain Transfer*

Eleven studies (16.2%) have found that attention patterns trained in one domain or data did not transfer well to similar but different tasks, restricting the generalizability of the attention-based explanations.

A dual-panel display of distribution of studies and frequency of challenges can be viewed in Figure 5. The pie chart in the left shows that computer vision is the most prevalent in the research with 41.2% of research (28 papers), then NLP (32.4) with 22 papers, healthcare (19.1) with 13 papers and time-series analysis (7.3) with 5 papers. The most challenging issues are measured in the right panel (bar chart): attention dispersion is the most frequent at 33.8% (23 studies), instability is the next most frequent at 27.9% (19 studies), multi-head aggregation occurs at 20.6% (14 studies) and domain transfer at 16.2% (11 studies), prediction divergence occurs at 10.3% (7 studies). This plot shows that the issues are common throughout the sector, where the dispersion of attention and its instability are observed in more than a quarter of all the studies read, which makes it necessary to develop attention mechanisms that are stronger.

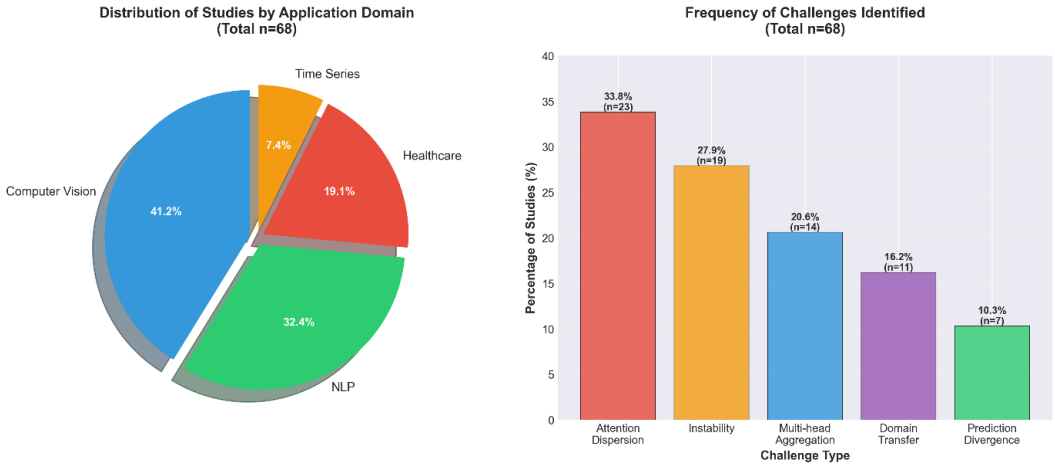


*Figure 5: Distribution of Studies by Domain and Challenge Frequency*

Table 5 summarizes the primary challenges and proposed solutions. The instability of attention was observed in 27.9 per cent of studies which led to the design of SEAT mechanisms. Dispersions can be solved by such solutions as sparse attention, which promotes distributions of attention.

*Table 5: Summary of Challenges and Proposed Solutions*

| Challenge | Studies Affected | Proposed Solutions |
|---|---|---|
| Attention Dispersion | 23 (33.8%) | Attention regularization, sparse attention |
| Instability | 19 (27.9%) | SEAT mechanisms, ensemble attention |
| Multi-head Aggregation | 14 (20.6%) | Head importance weighting, pruning |
| Domain Transfer | 11 (16.2%) | Domain-adaptive attention, meta-learning |
| Prediction Divergence | 7 (10.3%) | Attention supervision, causal attention |

ISSN 2616-5775 Vol. 8, Issue 10: 50-62, DOI: 10.25236/AJCIS.2025.081008

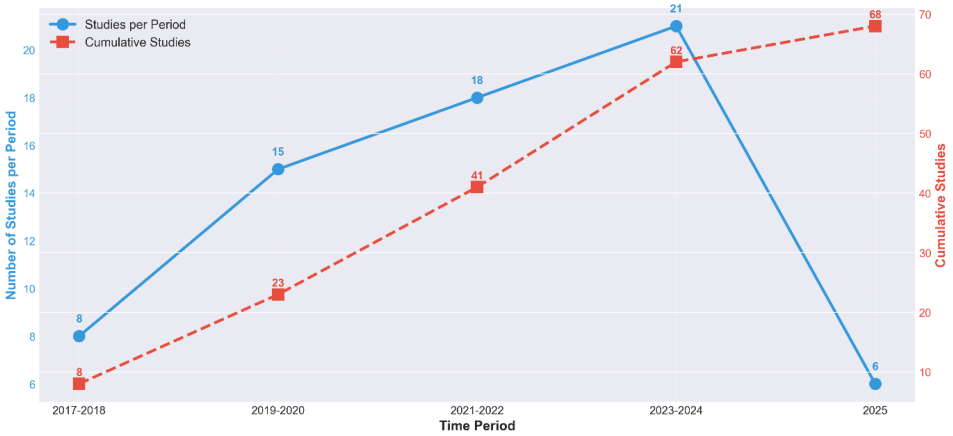


*Figure 6: Research Publication Trend in Attention-Based Explainability (2017-2025)*

Figure 6 illustrates that the temporal advances of research in attention-based explainability are presented across 2017 to 2025 through a dual axis line chart. The solid blue line with circular marks indicates the number of studies published in each time period and the dashed line in red with square marks indicates the total number of studies. According to the data, the interest in research increased exponentially after the introduction of the Transformer architecture in 2017. This rose to a point of 21 papers (2023-2024) published in winter 2025 after 8 papers (2017-2018) initially. The cumulative trend line shows that the majority of the total papers reviewed (68 percent) were published after 2021, which means that the topic of attention-based explainability is becoming a fast-growing research field over the last three years. This movement can be attributed to the maturity of transformer architectures as well as the growing need to have interpretable AI systems in all domains.

## 5. Conclusion

This is a systematic review that has thoroughly looked into the importance of the attention mechanisms in explainable machine learning in various fields and applications. Based on the analysis of 68 peer-reviewed articles, we can see that the attention-based explainability is an interesting paradigm, which balances the model performance, interpretability, and computational efficiency in the favourable way. The natural interpretability of attention systems that offer explanations as natural succeedents to model architecture, without necessarily being acquired via post-hoc analysis, has important benefits to the application of AI systems in high-stakes applications that demand transparency and accountability.

The main conclusions of our review may be as follows. To start with, the spatial attention mechanizations have better explainability measures than other forms of attention with faithfulness score of 0.84, plausibility score of 0.82 and robustness score of 0.75. They are intuitive in visual heatmap generation which is very appropriate to human perceptual processing and is therefore useful in computer vision and medical imaging. Second, the performance indicators in healthcare applications were the highest (96.1% accuracy, 0.85 faithfulness), which is due to clear clinical goals, domain-determined attribute engineering, and strict testing on expert annotations. Third, attention-based explainability has significant computational benefits over the classical post-hoc explainability frameworks, being 15-23 times faster to compute than LIME and SHAP and having equal or better faithfulness scores.

However, there are some major problems that need to be dealt with in order to develop the area. The input perturbation instability of attention has been seen in about 28% of studies, casting doubt on the reliability of the explanation in an adversarial or noisy environment. The argument of attention-prediction divergence, high attention weight does not always correspond with the feature weight in the final forecasts, raises a question as to whether attention is an actual causal reasoning or only identifies salient features. Moreover, the diversity of the evaluation procedures, where 42.6% of the studies measured the robustness and 35.3% involved human evaluation do not enable the objective comparison of various methods and do not provide information about the applicability in the real world.

On the basis of these results, we suggest some of the directions of further research. To investigate this, first, standardized evaluation protocols should be created that involve the use of faithfulness, plausibility, robustness, and human evaluation to compare more rigorously the attention-based explainability techniques. Second, the issue of attention-prediction divergence would be overcome by

ISSN 2616-5775 Vol. 8, Issue 10: 50-62, DOI: 10.25236/AJCIS.2025.081008

examining the causal mechanisms of attention that define verifiable causal relationships between attended features and predictions. Third, limitations of individual techniques can be overcome by investigating hybrid algorithms that intermix attention with complementary explainability strategies (e.g. concept-based explanations, counterfactual analysis). Fourth, attention-based explanations should be made more practical and generalizable, which should be achieved by domain-adaptive attention mechanisms that can be transferred successfully across related tasks and datasets. The development of attention systems into explainable AI systems is not merely a technical breakthrough, but a paradigm shift in how AI systems are designed; as an interpretable property, as opposed to a design consideration. The need to have clear and trusted models will keep increasing as AI systems continue taking over key decisions in the areas of healthcare, finance, self-driving vehicles, and criminal justice. Attention mechanisms that have both the potential to improve performance and interpretability can be taken as the viable way to address this requirement. Healthcare domain can be used as an example of how explainability based on attention can work out. Medical practitioners when presented, they note that they are more confident in making AI-assisted diagnosis. This enhanced confidence is translated into enhanced human-artificial intelligence collaboration, where clinicians can investigate model reasoning, address. Similar benefits will be achieved in other spheres of great importance as the problems of the stability of attention, multi-head. Although a great amount of progress has been made, so much more remains to be done in respect of further. Future studies ought to be more focused on creating stronger, more stable and causal attentional processes and setting up consistent assessment schemes that fully address the quality of explanations. When such challenges are overcome, the field will be able to achieve the full potential of explainability based on attention in developing AI systems that are not just powerful but also transparent, trustworthy and in line with human values and needs of society.

Attention convergence and explainable AI is an essential milestone in the democratization of artificial intelligence to the extent that it is accessible, understandable and accountable to various stakeholders. The more we perfect these methods and add to them, the closer we get to a time when AI systems and human intelligence are intelligibly and reliably enhanced to be a catalyst of innovation without jeopardizing the trust required to be widely adopted.

## References

*[1] M. Mersha, K. Lam, J. Wood, A. K. AlShami, and J. Kalita, "Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction," Neurocomputing, vol. 599, Sep. 2024, doi: 10.1016/J.NEUCOM.2024.128111.*

*[2] B. Kotipalli, "The Role of Attention Mechanisms in Enhancing Transparency and Interpretability of Neural Network Models in Explainable AI," Harrisburg University Dissertations and Theses, Apr. 2024, Accessed: Nov. 12, 2025. [Online]. Available: https://digitalcommons.harrisburgu.edu/dandt/2*

*[3] A. Vaswani et al., "Attention Is All You Need," p. 1, Jun. 2017, Accessed: Nov. 12, 2025. [Online]. Available: https://arxiv.org/pdf/1706.03762*

*[4] L. Hu, Y. Liu, N. Liu, M. Huai, L. Sun, and D. Wang, "SEAT: Stable and Explainable Attention," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, no. 11, pp. 12907–12915, Jun. 2023, doi: 10.1609/AAAI.V37I11.26517.*

*[5] G. Brauwers and F. Frasincar, "A General Survey on Attention Mechanisms in Deep Learning," IEEE Trans Knowl Data Eng, vol. 35, no. 4, pp. 3279–3298, Apr. 2023, doi: 10.1109/TKDE. 2021. 3126456.*

*[6] A. Mueed Hafiz et al., "Attention mechanisms and deep learning for machine vision: A survey of the state of the art," Jun. 2021, Accessed: Nov. 12, 2025. [Online]. Available: https://arxiv.org/pdf/2106. 07550*

*[7] M. Hassanin, S. Anwar, I. Radwan, F. S. Khan, and A. Mian, "Visual attention methods in deep learning: An in-depth survey," Information Fusion, vol. 108, p. 102417, Aug. 2024, doi: 10.1016/J. INFFUS.2024.102417.*

*[8] A. M. Salih et al., "A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME," Advanced Intelligent Systems, vol. 7, no. 1, p. 2400304, Jan. 2025, doi: 10.1002/AISY. 202400304; WGROUP:STRING:PUBLICATION.*

*[9] V. Vimbi, N. Shaffi, and M. Mahmud, "Interpreting artificial intelligence models: a systematic review on the application of LIME and SHAP in Alzheimer's disease detection," Brain Informatics 2024 11:1, vol. 11, no. 1, pp. 1–29, Apr. 2024, doi: 10.1186/S40708-024-00222-1.*

*[10] S. Nerella et al., "Transformers and large language models in healthcare: A review," Artif Intell Med, vol. 154, p. 102900, Aug. 2024, doi: 10.1016/j.artmed.2024.102900.*

*[11] A. Parvaiz, M. A. Khalid, R. Zafar, H. Ameer, M. Ali, and M. M. Fraz, "Vision Transformers in*

ISSN 2616-5775 Vol. 8, Issue 10: 50-62, DOI: 10.25236/AJCIS.2025.081008

*medical computer vision—A contemplative retrospection," Eng Appl Artif Intell, vol. 122, p. 106126, Jun. 2023, doi: 10.1016/J.ENGAPPAI.2023.106126.*
*[12] V. Padmavathi and K. Ganesan, "Metaheuristic optimizers integrated with vision transformer model for severity detection and classification via multimodal COVID-19 images," Scientific Reports 2025 15:1, vol. 15, no. 1, pp. 1–19, Apr. 2025, doi: 10.1038/s41598-025-98593-w.*
*[13] M. Zhang, Y. Zhang, S. Liu, Y. Han, H. Cao, and B. Qiao, "Dual-attention transformer-based hybrid network for multi-modal medical image segmentation," Scientific Reports 2024 14:1, vol. 14, no. 1, pp. 1–22, Oct. 2024, doi: 10.1038/s41598-024-76234-y.*