Engineering and Technology Journal e-ISSN: 2456-3358

Volume 11 Issue 04 April-2026, Page No.- 9447-9453

DOI: 10.47191/etj/v11i04.14 I.F. – 8.595

A Survey of Interpretable and Black-Box Deep Neural Networks

Ahmed Alhurdi

Department of Computer Science University of Science and Technology Aden, Yemen

ABSTRACT: This paper provides a more detailed overview of deep neural network (DNN) models with an emphasis on the potential for interpretability. The provided analysis categorizes the existing models based on interpretability (i.e., interpretable, post-hoc interpretable, or fully black-box models). The type of approach chosen for conducting the survey follows a well-structured methodology from several academic databases to comprehensively analyze the selected studies. The paper also analyzes explainability techniques such as SHAP and LIME, and investigates the trade-off in performance vs. complexity vs. interpretability with the existing models. The results of the analysis indicate that even though the black-box models have dominated in terms of peak accuracy, interpretable methods are improving comparably, achieving competitive results. Finally, this survey identifies three primary areas for future research that require attention: increased research related to model efficiency, transparency and scalability; lack of research addressing integration of explainability techniques into existing models; and unified low-complexity interpretability methods/models.

KEYWORDS: Deep Neural Network (DNN), CNN, Interpretable, Uninterpretable, LSTM, Black box, DRBFNN

I. INTRODUCTION

Artificial Intelligence (AI) is one of the most significant and rapidly growing fields in computer science, with the purpose of building intelligent systems capable of executing tasks that normally require human intelligence, for instance, reasoning, learning, insight, and decision-making. AI allows machines to adjust to dynamic environments, diagnose designs, and solve complicated problems competently through data-driven and knowledge-based approaches. According to Haenlein and Kaplan (2019), AI can be approximately (generally) defined as "a system's capacity or ability to appropriately interpret external data, to learn from such data, and to use those learnings to reach specific results and tasks through flexible adaptation [1]. Machine Learning (ML) is a subfield of AI that emphasizes emerging algorithms and models empowering computers to automatically learn and improve from experience without obvious programming. In machine learning systems, the led on a specific task improves as the system obtains more disclosure (detection) to data. Jordan and Mitchell (2015) define machine learning as "a scientific discipline interested in the question of how to construct computer programs that automatically optimize with knowledge." Machine learning pipelines involve data collection, preprocessing, feature extraction, model training, evaluation, and publishing or deployment, ensuring accurate and competent predictions or classifications [2]. Deep Learning (DL) is an advanced subfield of machine learning that uses multi-layered artificial neural networks to automatically learn hierarchical feature representations from data. Early layers in a deep network capture low-level features, while deeper layers represent more abstract patterns, making deep learning particularly effective for processing unstructured data such as images, audio, and text. As described by LeCun, Bengio, and Hinton (2015), deep learning "allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction." Core components of deep learning systems involve layers of neurons, activation functions, optimization algorithms, loss functions, and specialized architectures such as CNNs, RNNs/LSTMs, and Transformers [3].

explainable and Black-Box Models in deep learning. Deep learning models can be broadly split into explainable and black-box techniques. Interpretable techniques, such as linear regression and decision trees, enable users to easily comprehend how inputs affect outputs. In contrast, complicated deep neural networks are often treated as black boxes since their internal decision mechanisms are not directly observable. Therefore, interpretable techniques such as LIME (Local Interpretable Model-agnostic Explanations), SHAP (Shapley Additive Explanations), and Grad-CAM are employed to make these models more transparent. According to Molnar (2022), model interpretability is essential for ensuring trust, responsibility, and fairness in artificial intelligence applications [4].

Based on the above, the objective of this study is to analyze previous studies in the field of deep learning to determine research gaps, such as structure or architecture in interpretable and uninterpretable models, and massive

parameters used, to develop a novel algorithm that handles these gaps and reaches higher results in future work.

The main objective of this study is to provide a comprehensive and structured analysis of deep neural network models with respect to interpretability. Specifically, this survey aims to systematically classify deep learning models based on their level of interpretability, including intrinsic and post-hoc approaches, analyze the major explainability techniques such as SHAP, LIME, and Grad-CAM, and investigate the trade-off between model performance, complexity, and interpretability. Furthermore, this study seeks to identify key research gaps and challenges in the development of interpretable deep learning models, providing insights that can guide future research toward more efficient, transparent, and high-performance architectures.

Research in this area is commonly published in leading venues such as IEEE Transactions, ACM conferences, and high-impact journals in machine learning and artificial intelligence.

The main contribution of this paper is to provide a structured taxonomy and comparative analysis of interpretable and black-box deep neural networks, highlighting key challenges and research gaps.

This survey follows a structured methodology to ensure a comprehensive and unbiased selection of relevant studies. Multiple academic databases, including IEEE Xplore, Scopus, Web of Science, ACM Digital Library, and Google Scholar, were used to collect relevant literature. The search process was conducted using specific keywords such as "Deep Learning," "DNN," "Explainable AI," and "Interpretability." Inclusion criteria focused on peer-reviewed papers published between 2020 and 2025 that address interpretability or performance in deep neural networks. After an initial screening process, a total of 25 high-quality studies were selected for detailed analysis.

II. PRELIMINARIES

This section presents the fundamental concepts and definitions required to understand the proposed framework. Deep Learning is a subset of Machine Learning that enables models to learn directly from raw data without manual feature engineering [5]. In contrast, Deep Neural Networks are often considered black-box models due to their complex and non-transparent internal structure, making it difficult to interpret their decision-making process [6]. To address this limitation, several interpretability techniques have been introduced, such as Local Interpretable Model-Agnostic Explanations (LIME), Shapley Additive Explanations (SHAP), and Gradient-weighted Class Activation Mapping (Grad-CAM), which aim to provide explanations for model predictions and improve transparency [6].

III. TAXONOMY

The studies are classified into:

I- Intrinsic Interpretable Models

II Post-hoc Interpretability Models

III- Black-box Models

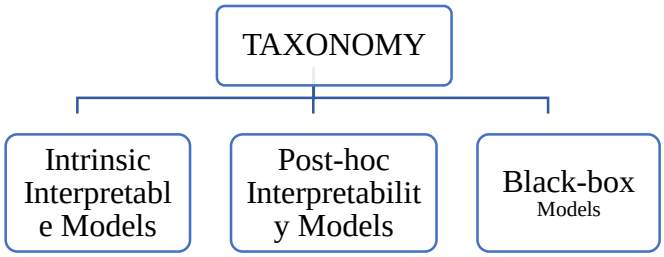

Figure1 Demonstrates Area Taxonomy

IV. LITERATURE REVIEW

Previous Studies (Structured Review)

This section demonstrates the structure review, related work, Interpretable Models, and black box models.

Interpretable Models

Several studies focused on designing interpretable models or applying explanation techniques. Studies such as [8], [9], and [13] introduced intrinsic interpretability by modifying neural network structures using regression coefficients, top-layer tracing, and minimum description length (MDL). These approaches improve transparency while maintaining competitive performance. Other studies [17], [21], and [23] applied feature attribution techniques such as DeepLIFT, sensitivity analysis, and DGA to quantify feature importance in prediction tasks. Post-hoc methods are widely used in recent works. Studies [28], [29], and [30] employed SHAP, Grad-CAM, and LIME to explain model predictions in healthcare and image analysis. These methods provide high interpretability while preserving model performance. Advanced approaches such as neural-symbolic models [26] and logic-rule extraction [27] attempt to convert neural network outputs into human-understandable rules.

Black-box Models

Several studies focused on improving performance without interpretability.

Optimization-based approaches such as [7] and [14] use Tabu Search and Gradient Descent Tunnelling to enhance model training. Model efficiency techniques such as pruning and feature selection are presented in [10] and [11], reducing complexity while maintaining accuracy. Ensemble and hybrid models are widely adopted. Studies such as [15] and [16] combine multiple neural networks or integrate decision trees with DNNs to improve performance. Deep architectures such as [18], [19], and [20] apply complex models (e.g., DBPNN, BDLSTM, FFDNN) for tasks like image classification and DDoS detection, achieving high accuracy. Comparative studies such as [22] highlight that deep learning outperforms traditional methods when large datasets are available.

V. ANALYSIS AND DISCUSSION

Analysis of Previous Studies

this section also has more of a critical evaluation of previous research, by looking at a particular set of studies, and then

sorting those studies according to their methods of interpretability used, and the type of models used in that set of studies. As summarized above in the Introduction section, studies of machine learning interpretable models have largely been conducted using two primary methods: Intrinsic Interpretability Models and Structural Models. The Intrinsic Interpretation Model category includes models such as LocalGLMnet and IFFNN that contain an inherent method of interpretability within the model architecture. These models employ mechanism such as regression coefficients (i.e., creating a line through all your data to show the direction of the relationship) or by tracing the features used in the top layer of the model to determine how a prediction was made. They perform very well, with many of these models achieving very high accuracy (90% or greater) for many tasks like classification of observations with tabular data, where the relationships of each of the relevant variables used to create each observation are clear and structured. The major benefit of these types of models compared to Structural Models is that they do not require an additional source of information for explanation since they are built in a manner that will be inherently transparent to their usage. A major disadvantage is that, because they are less scalable than Structural models, using Intrinsic Interpretation Models with more complex types of data (for example, using images or time series) produces less accurate results. Post-hoc Interpretability Models, which utilize post-training explanation techniques, are the second of three approaches discussed. SHAP, LIME, Grad-CAM and DeepLIFT are all common methods for interpreting the predictions of complex, deep learning models. Examples of using SHAP include predicting diabetes onset; Grad-CAM is frequently used in medical imaging to specify regions of interest on the image; and DeepLIFT is used for biological data analysis. Each of these approaches provides a highly predictive alternative with flexible and model-agnostic explanations, making them applicable throughout various fields, including healthcare and computer vision. Nevertheless, the results obtained with such methods will be approximate, rather than exact, and will add computational costs to the original model. Conversely, black-box models place a strong emphasis on developing complex architectures and advanced training methods that ultimately maximize the predictive performance of the model. These methods of training include techniques such as context decomposition, transfer learning, and creating an ensemble model (TabM, for example) by combining multiple models into a single model, as well as using optimization methods such as Tabu Search and GDT to help reduce the time needed to develop the model. Black-box models produce prediction accuracy (compared to all other models) that is often between 95% and 99% and can successfully predict outcomes based on various types of complex data sets. Additionally, the predictive ability of black-box models can be augmented by the use of ensemble or hybrid approaches. Despite being very accurate, black-box models have several significant disadvantages. First, they lack interpretability (i.e., transparency), which makes it difficult for decision-makers to understand what factors influence a specific prediction. Second, they are computationally complex to develop, and third, black-box models require a large number of tuning parameters. Therefore, black-box models cannot provide decision-makers with an adequate understanding of what information was included in the prediction-making process.

Table I presents a comparative analysis of interpretable, post-hoc, and black-box models across multiple dimensions, including performance, interpretability, and computational complexity. As observed, black-box models achieve the highest accuracy; however, they lack transparency. In contrast, intrinsic interpretable models provide high explainability with slightly lower performance. Post-hoc approaches offer a balanced solution by maintaining performance while providing flexible explanations.

Table I COMPARATIVE ANALYSIS OF INTERPRETABLE AND BLACK-BOX MODELS

Aspect	Intrinsic Interpretable Models	Post-hoc Interpretability Models	Black-box Models
Interpretability	High (built-in transparency)	Medium to High (explanation added externally)	Low (no inherent interpretability)
Performance	Moderate to High (≈90%+)	High (comparable to black-box)	Very High (≈95%–99%)
Model Complexity	Low to Moderate	High (due to base model + explanation tools)	Very High
Scalability	Limited for complex data	Good scalability	Excellent scalability
Data Suitability	Best for tabular/structured data	Works with all data types	Best for large-scale and complex data
Explanation Type	Intrinsic (e.g., coefficients, feature tracing)	Post-hoc (e.g., SHAP, LIME, Grad-CAM, DeepLIFT)	None (requires external tools)
Computational Cost	Low	Moderate to High	High
Flexibility	Limited (model-dependent)	Very high (model-agnostic)	Moderate (depends on architecture)

Use Cases	Finance, healthcare (structured data)	Healthcare, image analysis, NLP	Image classification, detection, and large-scale tasks

Comparative Analysis

A- Performance vs. Interpretability Trade-off

One of the major observations from the studies reviewed is that there is an inherent trade-off between predictive performance and model interpretability. For instance, black box models (particularly deep learning architectures) tend to produce higher peak accuracy when they have access to sufficient data to learn from, as they are comparatively better at analysing and learning from complex, non-linear relationships and types of data. In such cases, this increase in performance will often then reduce the level of transparency associated with the model and therefore increase the difficulty associated with the interpretation of decisions made by using the model. However, there have been significant advances made by interpretable models in attempting to reduce this gap, particularly for structured and tabular datasets, where the features relating to one another are relatively clear. This indicates that research is becoming increasingly focused on finding ways to achieve a balance between accuracy and explainability, without unduly compromising either attribute.

B- Impact of Model Complexity

The complexity of a machine learning system's model impacts both its effectiveness and interpretability. By increasing the number of layers, parameters, and components within a model, machine learning can better generalize patterns in data, leading to improved predictive accuracy; however, as a result of increased complexity in a model, it becomes less transparent and more difficult to understand and provide justification for its predictions. Conversely, lightweight or shallow models are typically more interpretable than their deeper counterparts, and in some cases, they have been shown to provide better generalization capabilities than other types of models, especially when trained with insufficiently sized data sets or moderately sized data sets. This emphasizes how important it is to choose an appropriate level of complexity for the type of problem being solved and the characteristics of the input data set.

C- Dataset Dependency

The type and structure of the dataset have a significant impact on which modeling approach is chosen to complete a project. For example, many image-based and unstructured data-related tasks will require the use of a black-box model (like a convolutional neural network) and subject to post-hoc interpretability techniques to provide a clear explanation of what happened. On the other hand, datasets containing tabular data (which is found in many industries, like healthcare and financial services) will typically lend themselves to the use of inherently interpretable models because they have a structured feature representation. This relationship between the dataset type and the model selected suggests that no single model will be the best option for all projects; instead, data will help determine the best piece of software.

D- Role of Hybrid Models

As more hybrid modelling methods are introduced, hybridization appears to be becoming a more common practice, which means integrating different methods to utilize their individual strengths. Examples of hybridization include convolutional neural networks integrated with recurrent neural networks and deep neural networks combined with gradient boosting methods. Typically, hybrid models achieve better results than either of the original model types by capturing different data patterns and relationships. The downside is that hybrid models typically require more complex model architecture, more computation, and generally have lower levels of explainability, which makes them more challenging to use due to the difficulties associated with balancing performance and explainability.

The above analysis and comparative evaluation highlight several limitations in existing approaches, which motivate the identification of key research gaps.

VI. KEY FINDINGS

Based on the comprehensive analysis of the reviewed studies, several important findings can be highlighted. First, there is a clear trade-off between interpretability and performance. Black-box models consistently achieve higher accuracy due to their ability to capture complex patterns, while interpretable models provide transparency but may struggle with highly complex data. Second, post-hoc interpretability techniques such as SHAP, LIME, and Grad- CAM have become the most widely adopted solutions, as they allow complex models to remain accurate while providing approximate explanations. Third, model complexity plays a critical role in both performance and interpretability. Increasing the number of layers and parameters improves accuracy but significantly reduces transparency and increases computational cost. Fourth, the effectiveness of different models strongly depends on the dataset type. Interpretable models perform well on structured and tabular data, whereas black-box models dominate in unstructured domains such as images and text. Finally, hybrid and ensemble approaches are emerging as promising solutions, combining the strengths of multiple models to improve performance. However, they further increase system complexity and reduce interpretability.

VII. OPEN PROBLEM

The literature review shows that there are several major gaps in our research that need to be explored further. First, most existing models show no unified model that can adequately balance interpretability with high predictive performance. Current methods tend to favour either one over the other, with little regard for those that may be equally important. Second, many of the proposed models have a very high computational complexity, which limits their use in practice and severely limits their use in real-time or when resources are limited. Third, many of the proposed models do not scale well; thus, many models will not be effective in either a large-scale dataset or when applied to high-dimensional datasets. Finally, there are no widely accepted evaluation standards for

measuring interpretability, making it difficult to measure and compare the ability of different models to explain their results consistently. These gaps demonstrate a need for future research to create more efficient, scalable, and balanced frameworks that will allow for accurate prediction and also for more reliable and standardised explanations for the predictions.

One of the most critical open problems is the lack of unified models that balance interpretability and performance without increasing computational complexity.

VIII. PROPOSED CONTRIBUTION

The purpose of the current study is to develop a new conceptual framework to help account for and fill in the gaps and problems with current models of machine learning by providing a solution that balances predictions with interpretable results and resource expenditure. As shown in Figure 2 below, the framework integrates "interpretability" directly into deep learning models/architectures, thus providing for making/explaining decisions in a more transparent and explainable manner, without having to drop model accuracy below acceptable levels. The method uses a combination of creating hybrid designs for the construction of interpretive aspects of the hybrid deep learning models and using a machine learning architecture that produces optimized performance characteristics. This combination of the hybrid design creates high transparency while also achieving extreme amounts of efficiency when it comes to extracting complex and non-linear behaviours from the data. In addition, this new approach focuses on simplifying the model (low number of model parameters and architecture size) so that it is scalable to real-world situations and does not require high computing resources, making it viable for use in real-world applications and environments where resources are constrained. Finally, the framework incorporates hybrid learning techniques to take advantage of using both black-box or non-interpretable models along with interpretable models to provide for improved generalisation performance while achieving a meaningful level of explainable predictions. The proposed framework will consist of the following six major components, as illustrated in Figure 1: (1) attribute/pre-feature data cleansing; (2) hybrid model generation containing an embedded interpretive aspect; (3) generation of evidence for the interpretation of the model (explanation); (4) the production of dynamic weights; (5) the fusion of evidence; and (6) the predictive output based upon the totality of the evidence.

Overall, the proposed framework addresses the key limitations identified in existing approaches, particularly the trade-off between interpretability and performance. It provides a unified and scalable solution that advances the development of efficient and transparent deep learning models for practical applications.

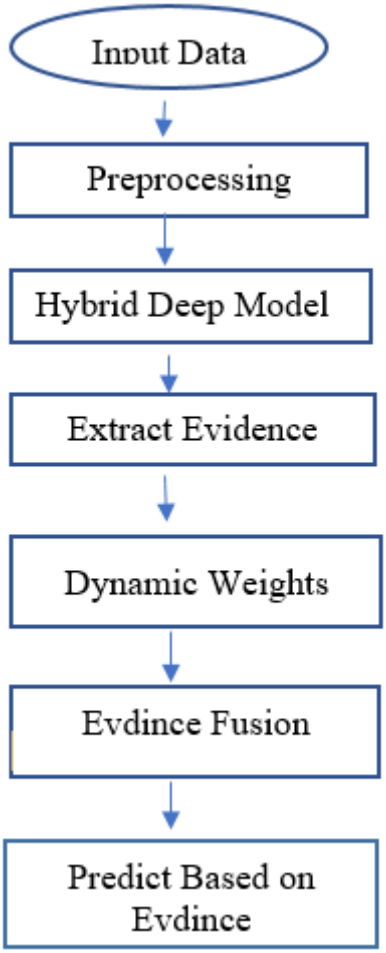

Figure 2 The proposed framework

IX. CONCLUSION

The current study provided an extensive overview of deep neural network (DNN) models with an emphasis on interpretability. This was accomplished through a structured investigation of intrinsic interpretable models, post-hoc explanation techniques, and black-box models. It revealed that there is still a trade-off between predictive accuracy, model complexity, and explainability as you would expect to see in any system. The study concluded that black-box models continue to provide better predictive accuracy than either intrinsic interpretable models or the post-hoc explanation techniques; however, the growth and evolution of post-hoc explanation techniques have led to an improvement in their ability to produce comparable performance with respect to various application types. The study raised numerous challenges regarding the scalability, computational efficiency, and the absence of standardised metrics for evaluating interpretability, while also identifying several gaps in the resource landscape that need to be addressed. Finally, the study provides some foundational research questions and suggestions that could assist future researchers in developing unified frameworks to provide an understanding of how to incorporate effective ways of understanding interpretability into high-performance models, improvements to the design of post-hoc explanation techniques, and more efficient/scale architectures.

In general terms, this document contributes to the body of knowledge by systematically categorizing previously

established methods, evaluating the individual attributes of each method and its associated strengths and weaknesses, then explaining how these findings may facilitate the construction of successful deep learning models through establishing a framework for constructing more interpretative, optimized, and dependable AI algorithms using DL technology, and finally pointing out the necessity for further studies aimed at bridging this gap in AI systems by combining the need for interpretability with that of achieving maximum performance.

X. FUTURE WORK

Future research directions include:

I- Designing hybrid interpretable architectures

II- Reducing model complexity

III- Improving scalability and efficiency

IV- Developing standardized XAI benchmarks

REFERENCES

1. M. Haenlein and A. Kaplan, "A brief history of artificial intelligence: On the past, present, and future of artificial intelligence," *California management review*, vol. 61, no. 4, pp. 5-14, 2019.
2. M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255-260, 2015.
3. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436-444, 2015.
4. C. Molnar, *Interpretable machine learning*. Lulu. com, 2020.
5. I. El Naqa and M. J. Murphy, "What are machine and deep learning? " in *Machine and deep learning in oncology, medical physics and radiology*: Springer, 2022, pp. 3-15.
6. X. Li *et al.*, "Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond," *Knowledge and Information Systems*, vol. 64, no. 12, pp. 3197-3234, 2022.
7. T. K. Gupta and K. Raza, "Optimizing deep feedforward neural network architecture: A tabu search-based approach," *Neural Processing Letters*, vol. 51, no. 3, pp. 2855-2870, 2020.
8. R. Richman and M. V. Wüthrich, "LocalGLMnet: interpretable deep learning for tabular data," *Scandinavian Actuarial Journal*, vol. 2023, no. 1, pp. 71-95, 2023.
9. M. Q. Li, B. C. Fung, and A. Abusitta, "On the effectiveness of interpretable feedforward neural network," in *2022 International Joint Conference on Neural Networks (IJCNN)*, 2022: IEEE, pp. 1-8.
10. L. Balderas, M. Lastra, and J. M. Benítez, "Optimizing dense feed-forward neural networks," *Neural Networks*, vol. 171, pp. 229-241, 2024.
11. Z. Liu *et al.*, "Ffsplit: Split feed-forward network for optimizing accuracy-efficiency trade-off in language model inference," *arXiv preprint arXiv:2401.04044*, 2024.
12. A. Charalampopoulos, N. Chatzis, F. Ntoulas-Panagiotopoulos, C. Papaioannou, and A. Potamianos, "Enhancing Fast Feed Forward Networks with Load Balancing and a Master Leaf Node," *arXiv preprint arXiv:2405.16836*, 2024.
13. S. Chatterjee and T. Sudijono, "Neural networks generalize on low complexity data," *arXiv preprint arXiv:2409.12446*, 2024.
14. B. Deng and L. Heath, "Toward Errorless Training ImageNet-1k," *arXiv preprint arXiv:2508.04941*, 2025.
15. Y. Gorishniy, A. Kotelnikov, and A. Babenko, "Tabm: Advancing tabular deep learning with parameter-efficient ensembling," *arXiv preprint arXiv:2410.24210*, 2024.
16. J. Yan, J. Chen, Q. Wang, D. Z. Chen, and J. Wu, "Team up gbdts and dnns: Advancing efficient and effective tabular prediction with tree-hybrid mlps," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024, pp. 3679-3689.
17. J. Martorell-Marugán *et al.*, "Explainable deep neural networks for predicting sample phenotypes from single-cell transcriptomics," *Briefings in Bioinformatics*, vol. 26, no. 1, p. bbae673, 2025.
18. J. S. Jeya I, "Deep Learning based Feed Forward Neural Network Models for Hyperspectral Image Classification," *The Open Biomedical Engineering Journal*, vol. 18, no. 1, 2024.
19. A. L. Yaser, H. M. Mousa, and M. Hussein, "Improved DDoS detection utilizing deep neural networks and feedforward neural networks as autoencoder," *Future Internet*, vol. 14, no. 8, p. 240, 2022.
20. P. Singh, S. K. Borgohain, A. K. Sarkar, J. Kumar, and L. D. Sharma, "Feed-Forward Deep Neural Network (FFDNN)-Based Deep Features for Static Malware Detection," *International Journal of Intelligent Systems*, vol. 2023, no. 1, p. 9544481, 2023.
21. S. Häffner, M. Hofer, M. Nagl, and J. Walterskirchen, "Introducing an interpretable deep learning approach to domain-specific dictionary creation: A use case for conflict prediction," *Political Analysis*, vol. 31, no. 4, pp. 481-499, 2023.
22. D. Tran and A. W. Tham, "Accuracy comparison between feedforward neural network, support vector machine and boosting ensembles for financial risk evaluation," *Journal of Risk and Financial Management*, vol. 18, no. 4, p. 215, 2025.
23. G. Jeon, H. Jeong, and J. Choi, "Distilled gradient aggregation: Purify features for input attribution in the deep neural network," *Advances in Neural Information Processing Systems*, vol. 35, pp. 26478-26491, 2022.

24. A. Eslamian and Q. Cheng, "TabMixer: advancing tabular data analysis with an enhanced MLP-mixer approach," *Pattern Analysis and Applications*, vol. 28, no. 2, pp. 1-17, 2025.
25. O. Dağlı, M. O. Kaya, and C. Eyüpoğlu, "Feed-Forward Neural Network-Based Prediction of Flutter Speed of 3 DOF 2D Wing Model," *Journal of Vibration Engineering & Technologies*, vol. 13, no. 1, p. 47, 2025.
26. S.-I. Jang, M. J. Girard, and A. H. Thiery, "Explainable and interpretable diabetic retinopathy classification based on neural-symbolic learning," *arXiv preprint arXiv:2204.00624*, 2022.
27. C. Geng, X. Xu, A. Xing, Z. Zhao, and X. Si, "From Neural Representations to Interpretable Logic Rules," *arXiv preprint arXiv:2501.08281*, 2025.
28. D. S. R. Kawale1, P. K. , and D. S. Holyachi, "Deep Neural Network Approach for Early-Stage Diabetes Risk Prediction using Hybrid SMOTE-ENN and GAN with SHAP-Based Feature Explanations " *Journal of Neonatal Surgery* vol. Vol. 14, Issue 31s (2025) no. ISSN(Online): 2226-0439 pp. 861-874, 2025.
29. M. Di Giammarco *et al.*, "Explainable Deep Learning for Breast Cancer Classification and Localization," *ACM Transactions on Computing for Healthcare*, vol. 6, no. 1, pp. 1-18, 2025.
30. K. Iftikhar, N. Javaid, I. Ahmed, and N. Alrajeh, "A Novel Explainable Deep Learning Framework for Accurate Diabetes Mellitus Prediction," *Applied Sciences*, vol. 15, no. 16, p. 9162, 2025.