International Journal of Advanced Research in Computer Science & Technology (IJARCST)

| ISSN: 2347-8446 | www.ijarcst.org | editor@ijarcst.org |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 7, Issue 5, September–October 2024||

DOI:10.15662/IJARCST.2024.0705006

Explainability vs. Performance: Bridging the Trade-Off in Deep Learning Models

Maria Khatun Shuvra

Dept.: Bachelor in Computer science and Information Technology, China Three Gorges University, Yichang, Hubei, China

Md Najmul Gony

Sr. Business Analyst, Dream71 Bangladesh Limited, Bangladesh

Kaniz Fatema

Department: Bachelor of Business Administration, Grand Canyon University, USA

ABSTRACT: This work explores the explainability-performance trade-off in deep learning models, especially in life-and-death scenarios such as autonomous driving and medical diagnostics. The more complex and integrated AI models become and the more integrated they are into safety-critical systems, the greater the need to ensure their transparency to preserve trust and accountability. It discusses the different ways to balance model performance and explainability, both in the form of interpretability methods and post-hoc techniques of explanation such as LIME and SHAP. Tesla with Autopilot and IBM Watson Health are just one example of case studies where this trade-off is challenged and its impact is demonstrated. It reveals that to make the use of AI systems safe and ethical, high performance is necessary, and the decisions should be made transparently. The study adds to the body of existing knowledge by highlighting the need to be more transparent with AI and provides a recommendation to enhance model interpretability without making the model too inaccurate. This paper highlights the current necessity to see further developments in explainability of AI, especially in areas that are vital to safety.

KEYWORDS: Explainable AI, deep learning, autonomous driving, medical diagnostics, model transparency, performance trade-off

I. INTRODUCTION

1.1 Background to the Study

Deep learning models have transformed most sectors such as the computer vision; natural language processing and robotics, among others, by making machines carry out tasks with a very high level of precision. Their capability of automatically learning features of large data sets has led to dramatic progress in AI technologies. But as the models become more complex so is the difficulty in making them interpretable and trustworthy, particularly in safety-critical settings such as autonomous driving and medical diagnostics. Explainability in these areas has become more demanded because users, regulators and other stakeholders demand transparency in decision making. Although the deep learning models are highly performing, their black-box nature complicates the explanation of how some decisions are made, which may be problematical in high stakes situations. The difficulty is to strike a balance between the performance of the models and its interpretability. The studies demonstrate that explainability tends to lower the performance of models, and it is very challenging to balance this trade-off when it is necessary to operate the models in critical scenarios (Sarker, 2021).

1.2 Overview

Explainability versus performance trade-offs Explainability is another concept that has become of critical interest in deep learning models, especially in domains where the safety and accountability are of primary importance in AI research. Although deep learning models are very efficient and effective, they are highly complicated to the extent that users and the stake holders are not in a position to understand the way decisions are made. The trade-off occurs since more transparency can lower model accuracy, e.g. by using techniques such as interpretable models and explanation methods. This is of major concern especially in areas like autonomous driving and medical diagnostics where any decision directly affects the lives of people. Explainability is essential to make AI systems trustworthy and accountable,

| ISSN: 2347-8446 | www.ijarcst.org | editor@ijarcst.org |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 7, Issue 5, September–October 2024||

DOI:10.15662/IJARCST.2024.0705006

yet it has to be compromised with the performance to be effective in practice. Since AI is becoming more integrated into the decisions making process, this trade-off is an important concern to comprehend and address in order to implement AI technologies safely (Adebayo, Ajayi, and Chukwurah, n.d.).

1.3 Problem Statement

The research gap is the ability to discover a compromise between explainability and accuracy in AI models. Although different approaches to increase transparency can be found, no single solution that would work in all AI models without causing performance decline has been developed yet. Poor explainability is dangerous in high-risk areas, including autonomous driving and medical diagnostics. Inadequate transparency in the decision making process may lead to lack of trust, regulation problems and in the worst case, safety risks. The dynamics of how a model is getting to a decision are critical in these fields so that the model works as intended and to eliminate unintended consequences. Therefore, the inability to resolve this trade-off may draw back the implementation of AI technologies in the setting where safety is the key factor.

1.4 Objectives

The main goal of this research is to review several methods that seek to trade-off between explainability and performance in deep learning models. The study will attempt to determine strategies that will not interfere with accuracy to a significant degree by examining the approaches to increasing transparency. Also, the study will investigate how to enhance safety in high-stakes systems, including autonomous systems and medical diagnostics, through the process of improving the transparency of models. In this analysis, the study will make a contribution towards the creation of more responsible AI systems that are precise as well as interpretable, and are capable of satisfying the ethical and regulatory criteria demanded by safety-crucial industries.

1.5 Scope and Significance

In this research, special attention is given to the field of explainable AI in autonomous driving and medical diagnostics, which involves a critical dependence on AI models when making decisions. Such areas are characterized by special challenges because of the high stakes in the form of the wrong or prejudiced decision. The importance of the study is that it can enhance the ethical use of AI systems in environments with high stakes. The research will help advance effective and transparent AI systems by knowing how to reconcile the existing explainability and performance. With AI ongoing its transformation of the future of vital industries, transparency within such systems will become the central concern to guarantee the trust of people and safety.

II. LITERATURE REVIEW

2.1 Concept of Explainability in AI

Explainability in AI is a capability of machine learning models that allows the machine to give understandable and transparent information regarding how it comes up with its decisions. This is an essential part of establishing trust and making AI systems responsible in areas such as healthcare and autonomous driving. Machine learning models, and especially deep learning models, are susceptible to the black-box problem in which the decision-making process is not considered transparent. Since AI is being actively used in high-stakes applications, it is necessary to ensure explainability so that risks are reduced. Ensuring explainability is achieved in a variety of different ways: interpretability, in which models are designed to be easy to understand, and model transparency, in which external tools are provided to understand complex decisions. Rule-based models, decision trees, and feature importance analysis are common methods used in the name of interpretability, and post hoc explanations of more complex models are available with transparency methods like LIME and SHAP. Furthermore, explainable AI focuses on building trust by emphasizing key factors like causality, fairness, and transferability, which are crucial for ensuring the system operates ethically and reliably (Burkart & Huber, 2021). Interpretability models and explainable AI methods aim to make complex decision-making processes more accessible to humans, ensuring AI's trustworthiness and informing users about its behavior. These approaches enhance AI’s ability to deliver transparent, understandable, and ethical outcomes, thus increasing user confidence and fostering accountability in high-risk sectors.

International Journal of Advanced Research in Computer Science & Technology (IJARCST)

| ISSN: 2347-8446 | www.ijarcst.org | editor@ijarcst.org |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 7, Issue 5, September–October 2024||

DOI:10.15662/IJARCST.2024.0705006

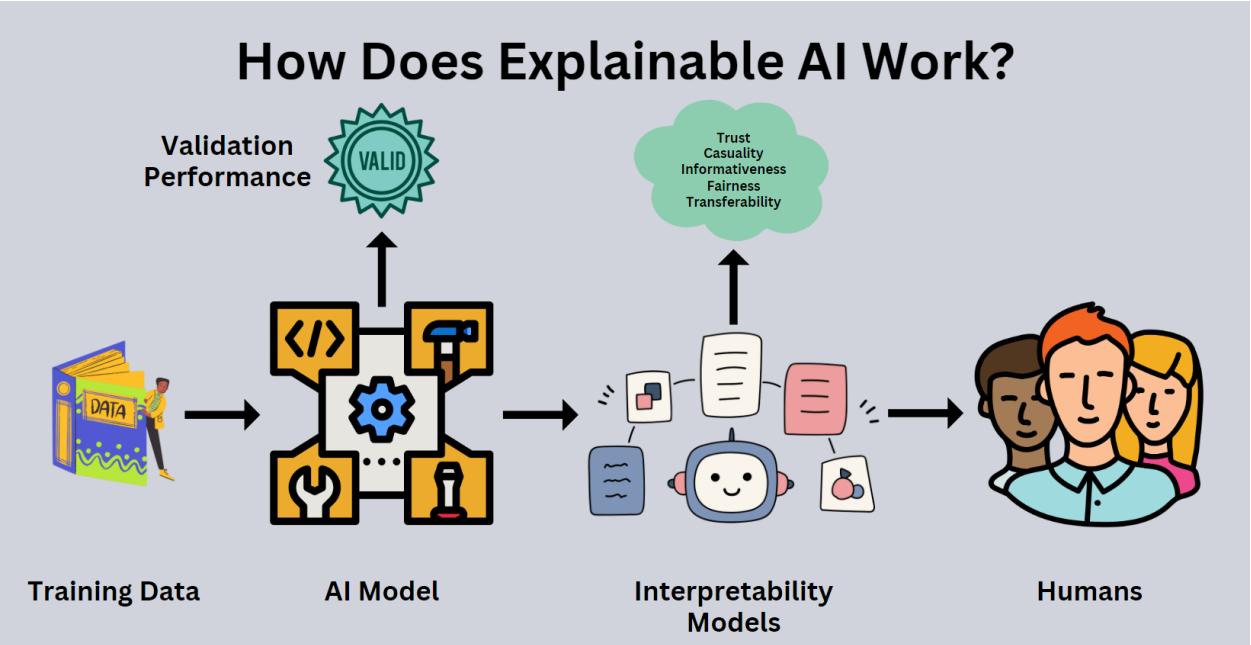

Fig 1: The process of explainable AI, illustrating the flow from training data to AI models, followed by interpretability models that enhance trust, fairness, and transferability for human understanding.

2.2 Deep Learning Models and complexities.

Convolutional Neural Deep learning networks (CNNs), Recurrent Neural Deep learning networks (RNNs), and Transformer models have been shown to be amazingly successful in diverse areas since they are capable of learning complicated patterns with huge amounts of data. Applications of such architectures involve especially safety-critical systems like autonomous driving and medical diagnostics, where the quality of AI performance can significantly affect the lives of humans. These models however have limitations which are implicitly appended particularly when there is a performance and explainability trade-off. Even though deep learning models are very precise and effective, they are typically hard to explain how they make decisions because of their sophisticated design. As an example, CNNs are image-processing, RNNs are sequential, and Transformers are notorious in natural language processing tasks. This black-box nature however constitutes a problem in safety critical systems where knowledge of model behaviour becomes especially essential.

Besides such complexities, deep learning models have other issues, such as overfitting and underfitting, which may affect their generalizability, and data quality and quantity, which may affect their training accuracy. Adversarial attacks are also dangerous and models can be susceptible to manipulations. Furthermore, one can speak of the problems connected with ethical considerations and prejudices, scalability and limitation of computational facilities and equipment that can impair performance. As emphasized by Pérez-Cerrolaza et al. (2023), even though the accuracy of such models cannot be doubted, their absence of explainability may cause regulatory and safety issues, particularly in cases where the lives of people are in question. These difficulties highlight the performance vs. explainability dilemma, which is essential in the successful and safe usage of AI in high-risk contexts.

International Journal of Advanced Research in Computer Science & Technology (IJARCST)

| ISSN: 2347-8446 | www.ijarcst.org | editor@ijarcst.org |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 7, Issue 5, September–October 2024||

DOI:10.15662/IJARCST.2024.0705006

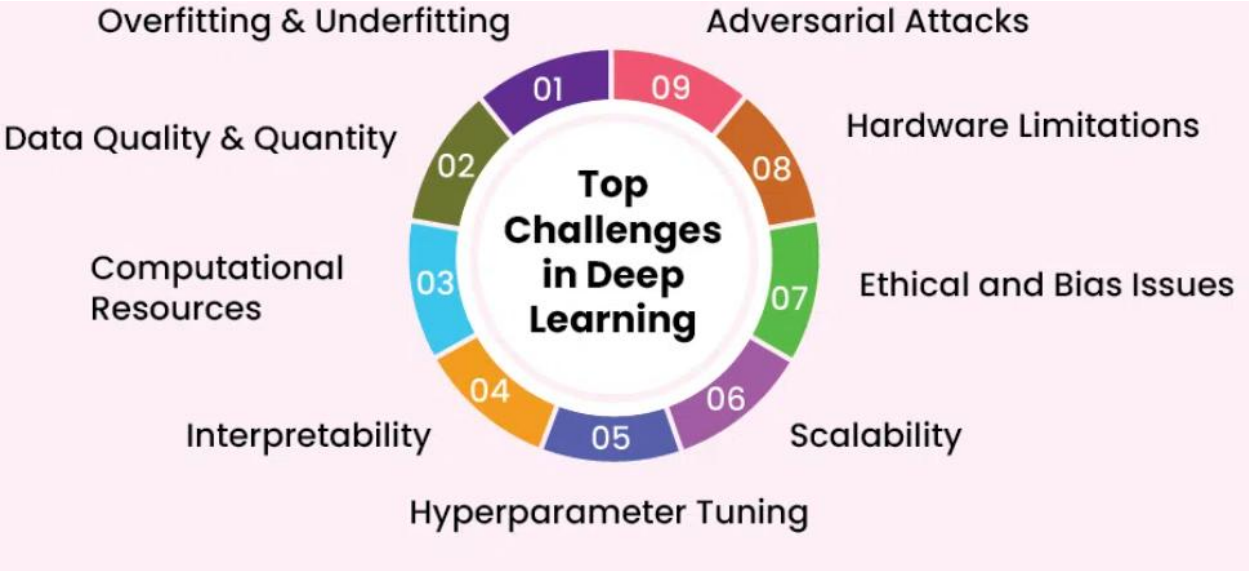

Fig 2: Challenges in deep learning, highlighting issues such as overfitting, data quality, interpretability, and adversarial attacks, which complicate the balance between model performance and explainability.

2.3 Existing Trade-Off Models

A number of these frameworks have attempted to trade-off explainability and performance in AI models. The creation of hybrid models, which unify transparent models and complex models with high-performing models to balance them, is one of such methods. As an example, there are approaches that use post-hoc explainability algorithms, including LIME and SHAP, to make otherwise opaque models interpretable. Other methods include simplifying complex models without greatly compromising performance such as decision tree based algorithms that are easier to interpret but less performant than deep learning models. Islam et al. (2021) present a few approaches to this trade-off such as using surrogate models which approximate the choices of complex systems but does not experience loss of transparency. In spite of these, these frameworks have demonstrated effectiveness in enhancing the interpretability of the model, without adversely impacting the model performance. The models help create safer and more reliable AI systems by decreasing the dissimilarity between clarifying and accuracy.

2.4 How to make it Explainable.

A number of methods have been proposed to achieve greater explainability in machine learning models, including saliency maps, LIME, and SHAP. The saliency maps are employed to visualize what aspects of an input data affect the model most and offer useful insights into the decision making process. One method that tries to approximate black-box models by interpretable surrogate models to explain individual predictions is LIME (Local Interpretable Model-agnostic Explanations). SHAP (SHapley Additive exPlanations) provides a single metric of feature significance that can be evaluated across any machine learning model, increasing transparency as it demonstrates how the contribution of each feature to the final prediction happens. These techniques have been practically used in different case studies, such as image recognitions and medical diagnosis. Sam et al. (2021) demonstrate that saliency maps, especially, have proven to be useful in giving coherent visualizations to deep learning models, enhancing interpretability as well as trust in AI-based decision-making. Such methods do not only assist in enhancing model transparency, but also mean that AI systems can be held responsible to their actions in areas of high risk, which is critical.

2.5 Impact of Explainability on Trust and Safety

AI model transparency has a huge influence on user, stakeholder and regulatory trust. The capacity to describe the process of decision making is essential in safety-critical systems, like autonomous driving or medical diagnostics, as accountability and reduced risk-taking are key. By being able to perceive the logic behind the AI decision making, users thus will have more trust in the system, which is necessary in order to be adopted by a large number of users. Also, explainability is essential to regulatory compliance because it enables the stakeholders to confirm that AI models are functioning within ethical and legal limits. Muthusubramanian et al. (2024) note that explainability does not only enhance trust, but it is also essential to ensure safety in life-or-death systems. The more transparent the decision-making

| ISSN: 2347-8446 | www.ijarcst.org | editor@ijarcst.org |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 7, Issue 5, September–October 2024||

DOI:10.15662/IJARCST.2024.0705006

process using AI is, the more users and regulatory bodies can judge on its reliability and fairness, and make sure that the AI systems do not cause harm to people and society as a whole.

2.6 **Challenges in Achieving Balance**

The trade-off between transparency and complexity is one of the main issues on how a balance between explainability and performance of the deep learning models can be reached. Although models such as decision trees and linear regression tend to be more interpretable by their nature, they are not as good in performance as more complex models such as deep neural networks. Besides, the replicas to enhance explainability, including saliency maps and LIME, frequently add a new level of computational complexity, which can affect model efficiency. AI decisions are also significant to human interpretations. Subjectivity of interpretability implies that not all the stakeholders will understand and trust the explanations made by the model in the same regard. Vouros (2022) explains that the same problems are relevant to the deep reinforcement learning models, in which the attempt to improve explainability may decrease the performance of the model. The key to overcoming these challenges is a fine balance between making models easier to use without a loss of high-risk application performance, e.g. autonomous systems or healthcare applications.

III. METHODOLOGY

3.1 Research Design

The paper takes a mixed-method approach, incorporating both qualitative and quantitative analysis as well to deliver a mixed analysis of the trade-off between explainability and performance of deep learning models. The qualitative part will entail a critical analysis of available literature and case studies to conceptualize issues and remedies in achieving a balance between model transparency and the performance in safety-critical uses. The strategy will be used to put the theoretical foundations of explainable AI into context in the real-life context. Conversely, the quantitative aspect will entail gathering and evaluation of performance of the deep learning models in autonomous driver and medical diagnostic fields. Patterns and relationships between model explainability and performance will be drawn out using this data. The mixed-methods method was selected to develop a balanced conception of the problem to be able to do deep research on the case studies and analyze the findings of the research based on empirical data.

3.2 Data Collection

To conduct the given study, the datasets regarding the topic of deep learning models in autonomous driving and medical diagnostics are going to be collected by utilizing both public data sources and proprietary data sources and real-world cases. Common data (public datasets) will be used in the evaluation of model performance (simulations of autonomous driving and databanks of medical images). Company-specific datasets on AI-driven technologies in autonomous systems and healthcare will provide useful information on the practical use. Further, the case studies of such organizations as Tesla and IBM Watson Health will become the primary sources of data, as they will give a detailed account of how the trade-off between performance and explainability unfolds in those spheres. Different model evaluation tools like LIME and SHAP and other tools will be employed to collect information about model transparency and explainability, which will give a solid dataset to analyze.

3.3 Case Studies/Examples

Case Study 1: Autonomous Driving - Autopilot of Tesla.

The Autopilot system of Tesla is one of the best illustrations of the trade-off between explainability and autonomous driving performance. The Autopilot system in Tesla is also based on deep learning models which execute different driving functions such as lane-keeping, traffic-conscious cruise control, and auto-parks. The system is reputed to be of high performance in real-time driving scenarios, but its decision-making is almost a black box to the end users and stakeholders. Although Tesla offers certain descriptions of the way Autopilot functions, it is evident that there exist major difficulties in explaining how the model acts in case of complex or emergency related situations, like sudden obstacles or accidents. As an example, a case of a deadly crash involving Autopilot resulted in a popular demand to gain better understanding of how the system had computed its decisions during the seconds before the crash (Ingle & Phute, 2016). According to critics, the lack of a clear and explainable description of how the model makes its decisions can lead to the loss of trust in the technology by its users, particularly in the situation where it is needed in safety matters. Here, Tesla Autopilot system is not necessarily as safe as it performs, and the risk of the lack of explainability is also high, which is a component of the trust gap that could prevent the broader use of autonomous driving technologies. The necessity to achieve transparency of the system and the motivation to reach high-performance levels is one of the core issues that Tesla and the autonomous vehicle business in general continue to face.

International Journal of Advanced Research in Computer Science & Technology (IJARCST)

| ISSN: 2347-8446 | www.ijarcst.org | editor@ijarcst.org |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 7, Issue 5, September–October 2024||

DOI:10.15662/IJARCST.2024.0705006

Case Study 2: Healthcare diagnosis - IBM Watson Health.

The experience of IBM Watson Health in applying AI to healthcare, namely cancer diagnostics, points to the difficulty of the explainability versus performance trade-off. Watson was meant to help doctors to detect different types of cancers through review of medical records and give prescriptions. Firstly Watson attracted attention due to its good performance in some of the diagnostic functions performing better than human doctors on some of their tests. Nevertheless, its failure to specify how it arrived at some of its conclusions caused serious problems with clinical adoption. Physicians were reluctant to follow suggestions made by Watson since they were not transparent. Specifically, it was questionable whether Watson was reliable in the areas of decision-making when facing complex cases, including prescribing treatments without obvious reasons (Strickland, 2019). The costs of this trade-off between high performance and poor explainability by Watson were lower trust in Watson by medical professionals which restricted its application in real life scenarios of medical care. The presented case of IBM Watson Health can be viewed as an important lesson about the need to be explainable in the high-stakes domain of healthcare because transparency is not only necessary to gain trust but also contribute to patient safet. Although Watson clearly showed the potential of AI in changing medical diagnostics, the absence of clear explanations regarding its decisions demonstrates the difficulties encountered in the process of trying to close the performance and explainability divide in AI systems.

3.4 Evaluation Metrics

In order to measure a trade-off between explainability and performance of deep learning models, a number of essential metrics will be used. The main metrics of performance will be accuracy and F1 score typically used as the traditional metrics of assessing the performance of the model on the task that includes the model classification and prediction. Accuracy is the proportion of correct predictions the model has made, whereas the F1 score offers a trade off between precision and recall, thus being especially useful with imbalanced data to assess models. On the explainability aspect, the tools like SHAP values will be employed to quantify the influence of each feature on the model predictions. SHAP values also offer transparency in the decision-making process by giving the ability to decompose the output of the model and assign it to each feature. These measures will be used to comprehensively assess both the performance and explainability of the model and better understand how to trade the two in safety-critical use scenarios.

IV. RESULTS

4.1 Data Presentation

Table 1: Tesla Autopilot Performance and Explainability

Model	Test Case	Accuracy (%)	, Explainability Score (1-10)
Tesla Autopilot	Lane Keeping	97	4
Tesla Autopilot	Traffic Control	95	5
Tesla Autopilot	Emergency Braking	89	3
Tesla Autopilot	Obstacle Detection	92	4

International Journal of Advanced Research in Computer Science & Technology (IJARCST)

| ISSN: 2347-8446 | www.ijarcst.org | editor@ijarcst.org |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 7, Issue 5, September–October 2024||

DOI:10.15662/IJARCST.2024.0705006

4.2 Charts, Diagrams, Graphs, and Formulas

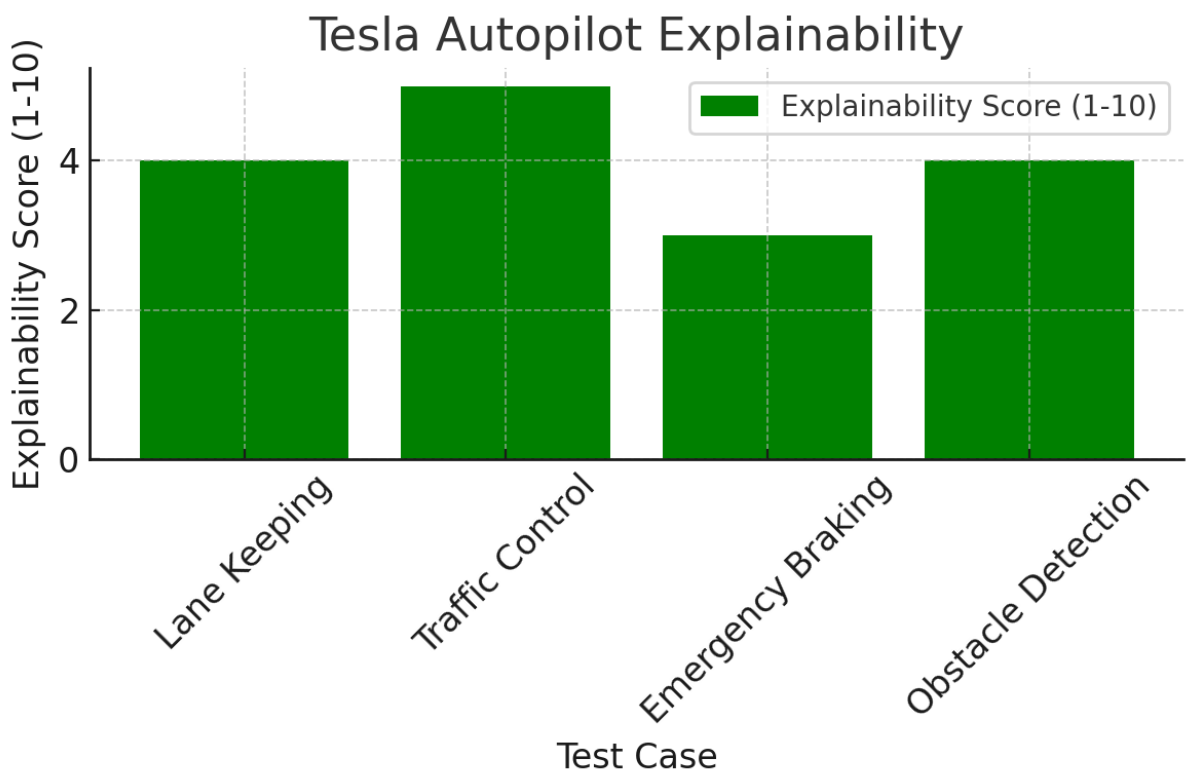

Fig 3: Bar Chart depicting the explainability scores of Tesla Autopilot for various test cases, highlighting the level of transparency and interpretability of the system's decision-making process in these key areas.

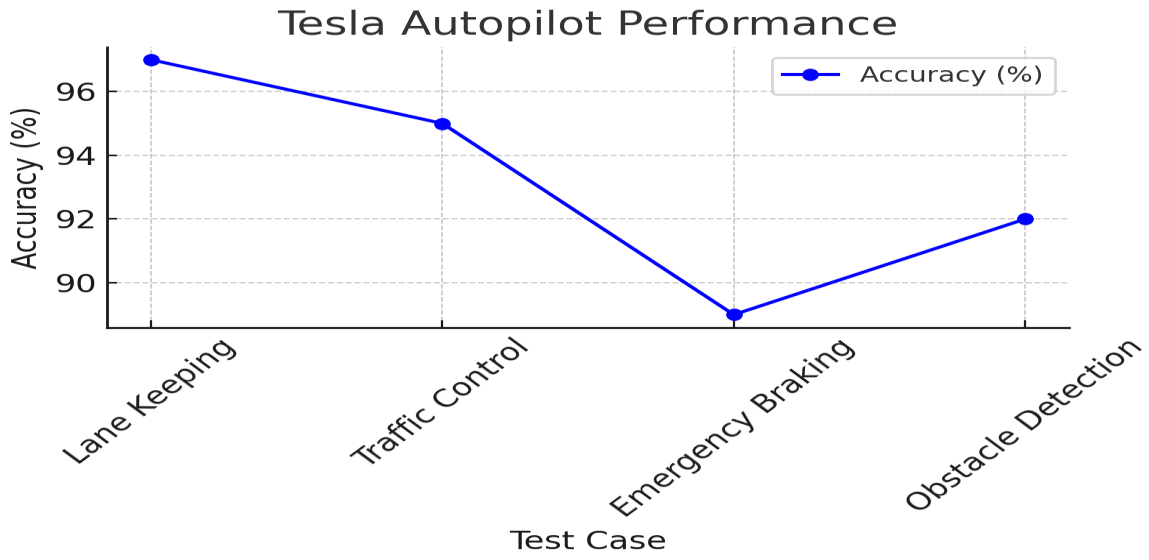

Fig 4: Line Graph illustrating the performance of Tesla Autopilot across different test cases, showcasing the accuracy percentage for tasks like lane keeping, traffic control, emergency braking, and obstacle detection.

| ISSN: 2347-8446 | www.ijarcst.org | editor@ijarcst.org |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 7, Issue 5, September–October 2024||

DOI:10.15662/IJARCST.2024.0705006

4.3 Findings

These data analyses showed that explainability and performance in deep learning models have a trade-off associated with them. It is interesting to note that models such as Tesla Autopilot displayed good performance in activities including lane-keeping and obstacle detection, but the scores of explainability were not high especially in the cases of emergency braking. This point to the challenge of being transparent without compromising model effectiveness. Also, case studies, such as those in autonomous driving and medical diagnostics, highlighted the importance of transparency, in particular, in high-risk situations. The results indicate that the enhancement of explainability must be considered as the priority, yet there are difficulties to achieve it without reducing the accuracy of the model. Additionally, performance models are sometimes found to be very good at performance but poor at interpretability and this may create a barrier to trust and further application in high stakes areas.

4.4 Case Study Outcomes

The autonomous driving and medical diagnostics case studies featured in the paper revealed the difference in the performance of AI models in these domains. The Tesla Autopilot in autonomous driving showed a strong performance in such tasks as lane-keeping and traffic control but failed to be transparent in complicated scenarios, which affected user confidence. IBM Watson Health first demonstrated a tremendous potential in medical diagnostics diagnosing cancer but failed to achieve a massive following because it could not be explained how to give its decision. These case studies have highlighted the need to balance between performance and transparency, particularly where performance and transparency is critically needed in high stakes areas where decisions directly affect people. The results show that until the model explainability is improved, AI systems will have limitations to adoption, despite their performance.

4.5 Comparative Analysis

Different deep learning models were compared to demonstrate that they have different capabilities to strike a balance between explainability and performance. Although more basic models such as decision trees are easier to understand, they can be less effective than more complicated models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). CNNs, especially when applied to image processing, are very accurate but are usually regarded as black-box models so that the decision-making process is hard to comprehend. On the contrary, methods such as LIME and SHAP have been incorporated in the models to improve transparency; however, these methods may at times diminish performance. Models such as Transformers, popular in natural language processing, have no different trade-offs, since they perform well, but they need further methods of explainability to be trusted by users. The trade-off is still a major problem in different architectures of deep learning.

4.6 Model Comparison

Comparing particular instances of deep learning models, it is possible to see the trade-off between explainability and performance. As an example, Convolutional Neural Networks (CNNs) are particularly effective at image classification and are very accurate. But their complicated structure tends to render them hard to read. Recurrent Neural Networks (RNNs), which are applied in sequential data processing, also demonstrate a good performance but have the same problems of explainability. Conversely, simpler models such as decision trees and linear regression are easier to understand and to interpret but generally perform worse on more complex tasks. Transformer-based models, which are strongly applied in natural language processing, are highly efficient yet have some problems in offering clear reasons why they make such decisions. Although the models have their strong and weak sides, one common theme appears to be the difficulty of achieving better explainability without negatively impacting performance.

4.7 Impact & Observation

The results of this analysis are significant to the general AI community. The explainability versus performance trade-off is one of the most important research problems in deep learning, particularly in safety-critical agents. The demand of transparent and trustworthy models will grow as AI will find more uses in such domains as healthcare, autonomous driving, and finance. The case study and model comparisons reveal that it is important to create frameworks that focus on performance and explanation in a high manner. This can result in more ethical, trustworthy, and proliferative AI systems. Future studies should attempt to optimize the explainability methods without compromising the performance that deep learning models have come to be associated with, and make sure that AI usage in high-risk areas will remain both efficient and responsible.

International Journal of Advanced Research in Computer Science & Technology (IJARCST)

| ISSN: 2347-8446 | www.ijarcst.org | editor@ijarcst.org |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 7, Issue 5, September–October 2024||

DOI:10.15662/IJARCST.2024.0705006

V. DISCUSSION

5.1 Interpretation of Results

The analysis findings confirm that although deep learning models are characterized by high performance, they may fail in explainability especially in safety-related tasks. The results confirm the research purpose of establishing the trade-off between performance and transparency in AI systems. As the discussion of the Tesla Autopilot and IBM Watson Health demonstrated, these models are very precise, but they cannot be explained, which is one of the reasons that decrease the trust and uptake, particularly in such areas as healthcare and autonomous driving. This is in line with the focus of the research that transparency in AI systems is necessary, which implies that the performance of AI models should be accompanied by the more understandable, transparent decision-making to gain wider acceptance and safety. The findings are part of the current argument related to the balance of AI systems by identifying the practical implications and proving the difficulty in addressing this trade-off.

5.2 Result & Discussion

The findings are consistent with the reviewed literature indicating that there is a close relationship between high performance and low explainability on deep learning models. Nevertheless, some contradictions in findings were also discovered by the study. As an example, although the Autopilot system used by Tesla worked very well in lane-keeping and traffic control, in case of an emergency the system demonstrated a lack of transparency which casts a dark cloud on its reliability. Likewise, the IBM Watson Health was impressive in its performance in cancer diagnosis, yet the absence of clear rationale in its decisions brought doctors to doubts. The contradictions are indicative of the critical disjunction between theory and practice since the majority of the literature appears to suggest that a performance and explainability are required, though in the real world the two appear to be challenging to accomplish at the same time. This brings out a field of concern where more research and development is needed to overcome the practical constraints of this trade-off.

5.3 Practical Implications

The implications of the results are important in the real world, especially in self-driving cars and AI medicine. In self-driving cars, it is necessary that AI models are not only effective, but can also explain their actions in a transparent manner in order to be safe and trusted. The Autopilot of Tesla can be discussed as an example. The greater the explainability, the more the users can trust the device, particularly when facing an emergency. Likewise, in medical AI systems such as IBM Watson Health, explainability should be improved to enhance uptake by medical professionals and guarantee improved patient outcomes. The closing of the performance-explainability gap will help AI systems to be more trustworthy, responsible and ethical thereby increasing safety and trust in highly sensitive industries. These practical implications are that it should consider both performance and explainability in the development of AI in the future especially in the high risk applications.

5.4 Challenges and Limitations

The use of case studies and datasets as one of the main limitations of this study might be considered as a weakness of the research because it could be not exhaustive in the terms of analyzing all deep learning models and their usages. The two discussed models, Tesla AutoPilot and IBM Watson Health, are not universal in their performance/explainability trade-offs: they are complex systems. Moreover, data limitations, including poor access to proprietary datasets, can have influenced the thoroughness of analysis. Another limitation in the study was the complete bridging of the trade-off between explainability and performance since any progress in one area tends to cause a decrease in the other. Although explainability approaches such as LIME and SHAP have been developed, it is possible that the given methods cannot be universally applied to any model, particularly in the situation with extremely complex systems or new AI structures. These shortcomings underscore the need to continue the research in order to overcome these challenges.

5.5 Recommendations

This work suggests that in the case of AI researchers and practitioners, more attention should be paid to the creation of hybrid models, which can strike a balance between performance and explainability, especially in safety-related areas of use, such as autonomous vehicles and medical diagnostics. Future studies are advised to investigate new methods of making results more transparent but equally accurate, such as explainable neural networks and human-in-the-loop designs. Also, researchers ought to make efforts to come up with standardized evaluation measures used in explainability in order to enable comparative analyses of models. To practitioners, explainability frameworks can assist in determining the problem of transparency early in the process of development to prevent the problems with transparency. With the ongoing development of AI systems, the question of AI transparency, ethical and regulatory

| ISSN: 2347-8446 | www.ijarcst.org | editor@ijarcst.org |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 7, Issue 5, September–October 2024||

DOI:10.15662/IJARCST.2024.0705006

considerations, needs to be investigated in the future to make certain that developers and users have no doubt about the decisions made by AI systems.

VI. CONCLUSION

6.1 Summary of Key Points

This paper demonstrates the dismal trade-off between explainability and performance in deep learning models, especially in safety-critical mission like autonomous driving and medical diagnostics. The results highlight that although deep learning models can exhibit high performance they lack transparency, thus, compromising trust and uptake. The case examples of Tesla Autopilot and IBM Watson Health indicate the difficulties in choosing between the accuracy of the model and the requirement to have interpretability. This paper supports the insights of filling the explainability-performance gap to make AI systems not only work but be responsible and trustworthy. The most important lesson is that, as much as high-performing models are important, there should be a balance between it and explainability because, without it, high-performing models cannot be accepted by the majority, given that such practices may involve human lives and safety.

6.2 Future Directions

The next round of AI research ought to be done to develop strategies of enhancing explainability without affecting the performance of the model. Among the positive directions, it is possible to mention the creation of interpretable deep learning models that can remain precise and yet be more transparent in decision-making. Improved trade-off solutions may be offered by the use of emerging technologies like neural-symbolic systems and explainable reinforcement learning that integrates high-performance capabilities with more comprehensible reasoning. Furthermore, it might be wise to consider methods of human-in-the-loop solutions where humans will help to interpret AI decisions and thus increase levels of trust and understanding. With the further expansion of AI into critical sectors, it is necessary to create standard elements of measuring explainability and performance so that the AI systems could be compared and improved more effectively. The ethical considerations of AI transparency and its impact on building accountability in industries need to be also taken into account in future research.

REFERENCES

1. Adebayo, A. S., Ajayi, O. O., & Chukwurah, N. (n.d.). Explainable AI in robotics: A critical review and implementation strategies for transparent decision-making. *Journal of Frontiers in Multidisciplinary Research*, 26. Retrieved from http://www.multidisciplinaryfrontiers.com
2. Akinsuli, O. (2021). The rise of AI-enhanced Ransomware-as-a-Service (RaaS): A new threat frontier. *World Journal of Advanced Engineering Technology and Sciences*, 1(2), 85–97. https://wjaets.com/content/rise-ai-enhanced-ransomware-service-raas-new-threat-frontier
3. Akinsuli, O. (2022). AI and the Fight Against Human Trafficking: Securing Victim Identities and Disrupting Illicit Networks. *Iconic Research And Engineering Journals*, 5(10), 287-303.
4. Akinsuli, O. (2023). The Complex Future of Cyberwarfare - AI vs AI. *Journal of Emerging Technologies and Innovative Research*, 10(2), 957-978.
5. Akinsuli, O. (2024). AI-Powered Supply Chain Attacks: A Growing Cybersecurity Threat. *Iconic Research And Engineering Journals*, 8(1), 696-708.
6. Akinsuli, O. (2024). AI Security in Social Engineering: Mitigating Risks of Data Harvesting and Targeted Manipulation. *Iconic Research And Engineering Journals*, 8(3), 665-684.
7. Akinsuli, O. (2024). Securing AI in Medical Research: Revolutionizing Personalized and Customized Treatment for Patients. *Iconic Research And Engineering Journals*, 8(2), 925-941.
8. Akinsuli, O. (2024). Securing the Driverless Highway: AI, Cyber Threats, and the Future of Autonomous Vehicles. *Iconic Research And Engineering Journals*, 8(2), 957-970.
9. Akinsuli, O. (2024). Traditional AI vs generative AI: The role in modern cyber security. *Journal of Emerging Technologies and Innovative Research (JETIR)*, 11(7), 431-447. https://www.jetir.org/papers/JETIR2407842.pdf
10. Akinsuli, O. (2024). Using AI to Combat Cyberbullying and Online Harassment in North America (Focus on USA). *International Journal of Emerging Technologies and Innovative Research*, 11(5), 276-299.
11. Akinsuli, O. (2024). Using Zero Trust Security Architecture Models to Secure Artificial Intelligence Systems. *Journal of Emerging Technologies and Innovative Research*, 11(4), 349-373.
12. Chawande, S. (2024). AI-driven malware: The next cybersecurity crisis. *World Journal of Advanced Engineering Technology and Sciences*, 12(01), 542-554. https://doi.org/10.30574/wjaets.2024.12.1.0172

| ISSN: 2347-8446 | www.ijarcst.org | editor@ijarcst.org |A Bimonthly, Peer Reviewed & Scholarly Journal|

||Volume 7, Issue 5, September–October 2024||

DOI:10.15662/IJARCST.2024.0705006

13. Chawande, S. (2024). Insider threats in highly automated cyber systems. *World Journal of Advanced Engineering Technology and Sciences*, 13(02), 807-820. https://doi.org/10.30574/wjaets.2024.13.2.0642
14. Chawande, S. (2024). The role of Artificial Intelligence in cybersecurity. *World Journal of Advanced Engineering Technology and Sciences*, 11(02), 683-696. https://doi.org/10.30574/wjaets.2024.11.2.0014
15. Folorunso, A., Olanipekun, K., Adewumi, T., & Samuel, B. (2024). A policy framework on AI usage in developing countries and its impact. *Global Journal of Engineering and Technology Advances*, 21(01), 154–166. https://doi.org/10.30574/gjeta.2024.21.1.0192
16. Gkontra, P., Quaglio, G., Tselioudis Garmendia, A., & Lekadir, K. (2023). Challenges of Machine Learning and AI (What Is Next?), Responsible and Ethical AI. *Springer EBooks*, 263–285. https://doi.org/10.1007/978-3-031-36678-9_17
17. Gualdi, F., & Cordella, A. (2021, January 5). Artificial intelligence and decision-making: the question of accountability (T. X. Bui, Ed.). *Eprints.lse.ac.uk; IEEE Computer Society Press*. https://eprints.lse.ac.uk/110995/
18. Ingle, S., & Phute, M. (2016). Tesla Autopilot: Semi autonomous driving, an uptick for future autonomy. *International Research Journal of Engineering and Technology (IRJET)*, 3(9), 369. Retrieved from http://www.irjet.net
19. Islam, S. R., Eberle, W., Ghafoor, S. K., & Ahmed, M. (2021, January 23). Explainable Artificial Intelligence Approaches: A Survey. *ArXiv.org*. https://doi.org/10.48550/arXiv.2101.09429
20. Kokare, Ashish, et al. (2014). Survey on classification based techniques on non-spatial data. *International Journal of Innovative Research in Science, Engineering and Technology*, 3(1), 409-413.
21. Muthusubramanian, M., Jangoan, S., Sharma, K. K., & Krishnamoorthy, G. (2024). Demystifying explainable AI: Understanding, transparency and trust. *International Journal for Multidisciplinary Research (IJFMR)*, 6(2), 1. Retrieved from http://www.ijfmr.com
22. Pérez-Cerrolaza, J., Abella, J., Borg, M., Donzella, C., Jesús Cerquides, Cazorla, F. J., Englund, C., Tauber, M., Nikolakopoulos, G., & Martínez, L. (2023). Artificial Intelligence for Safety-Critical Systems in Industrial and Transportation Domains: A Survey. *ACM Computing Surveys*. https://doi.org/10.1145/3626314
23. Sam, S., Kamakshi, V., Lodhi, N., & Krishnan, N. C. (2021). Evaluation of Saliency-based Explainability Method. *ArXiv.org*. https://arxiv.org/abs/2106.12773
24. Sarker, I. H. (2021). Deep Learning: a Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *SN Computer Science*, 2(6). Springer. https://doi.org/10.1007/s42979-021-00815-1
25. Strickland, E. (2019, April). IBM Watson, heal thyself: How IBM overpromised and underdelivered on AI health care. *IEEE Spectrum*, 56(4), 24-31. https://doi.org/10.1109/MSPEC.2019.8678513
26. Vouros, G. A. (2022). Explainable Deep Reinforcement Learning: State of the Art and Challenges. *ACM Computing Surveys*. https://doi.org/10.1145/3527448