International Journal for Modern Trends in Science and Technology
Volume 12, Issue 05, pages 30-36
ISSN: 2455-3778 online
Available online at: http://www.ijmtst.com/vol12issue05.html
DOI: https://doi.org/10.5281/zenodo.19613600

Explainable Deep Learning Model for Ethical and Transparent Decision Support Systems

Gajula Kiran, Gali Pavan Kumar Reddy, Gokavarapu Lokhanadha Durga Satish, Gorrela Saimurali, Gude Kiran Teja

Department of Computer Science and Engineering (AI & DS), Sir C R Reddy College of Engineering, Eluru, Andhra Pradesh, India

To Cite this Article
Gajula Kiran, Gali Pavan Kumar Reddy, Gokavarapu Lokhanadha Durga Satish, Gorrela Saimurali & Gude Kiran Teja (2026). Explainable Deep Learning Model for Ethical and Transparent Decision Support Systems. International Journal for Modern Trends in Science and Technology, 12(05), 30-36. https://doi.org/10.5281/zenodo.19613600

Article Info
Received: 28 March 2026; Revised: 24 April 2026; Accepted: 26 April 2026.

KEYWORDS

Explainable AI, Deep Learning, Decision Support Systems, SHAP, XAI, Transparency, Bias Detection, Neural Networks, Feature Importance, Ethical AI, TensorFlow, PyTorch.

ABSTRACT

Artificial Intelligence and deep learning models are widely used in decision support systems across domains such as healthcare, finance, and law. Although these models provide high predictive accuracy, they often operate as black boxes, making it difficult for users to understand how decisions are made. The lack of transparency raises concerns regarding trust, accountability, and ethical decision-making. This project proposes an Explainable Deep Learning Model for Ethical and Transparent Decision Support Systems. The primary objective is to improve the interpretability of deep learning predictions while maintaining high performance. The proposed model integrates Explainable Artificial Intelligence (XAI) techniques such as SHAP (SHapley Additive Explanations) and attention mechanisms to identify the most influential features contributing to each prediction. Fairness and bias detection techniques are incorporated to ensure ethical decision-making and reduce discrimination in automated predictions. The model is evaluated using real-world datasets from healthcare and financial transaction domains. Experimental results demonstrate that the proposed approach achieves competitive prediction accuracy while providing clear and interpretable explanations for model outputs. The results indicate that integrating explainability into deep learning models improves transparency, increases user trust, and supports responsible AI deployment. This work contributes to the development of ethical and transparent decision support systems that bridge the gap between complex machine learning models and human understanding.

1. INTRODUCTION

1.1 Background

In recent years, deep learning has emerged as a transformative technology across numerous domains, including healthcare, finance, autonomous systems, and legal decision-making. Its ability to automatically extract complex patterns from vast amounts of data has led to unprecedented advancements in predictive accuracy and automation. However, deep learning models are often criticised for their "black-box" nature, where the internal decision-making processes remain opaque and difficult to interpret.

This opacity poses significant challenges when deploying such models in decision support systems that require ethical considerations, transparency, and accountability. Decision support systems are widely used to assist professionals in making critical decisions based on data-driven insights. When the reasoning behind a model's prediction is unclear, users may find it difficult to trust or verify the system's outputs. As AI continues to integrate into real-world environments, the need for explainable and transparent AI systems has become increasingly important.

Regulatory bodies across the world have also begun emphasising the importance of transparency in automated decision-making systems. Data protection laws highlight the need for systems that can provide clear explanations for automated decisions affecting individuals. Another important challenge is the presence of bias in training data. If the data used to train models contains biases, the model may learn and reproduce these biases in its predictions, leading to unfair outcomes in domains such as hiring, lending, and criminal justice.

1.2 Importance of the Project

The project titled "Explainable Deep Learning Model for Ethical and Transparent Decision Support Systems" addresses an important issue at the intersection of artificial intelligence and societal responsibility. As AI systems become more integrated into everyday decision-making, ensuring that these systems operate transparently and ethically becomes essential. Explainable models enable users to understand how decisions are made, allowing stakeholders to verify the fairness and reliability of AI systems.

The importance of this project can be understood from several perspectives. First, ethical compliance: AI systems used in decision support must adhere to ethical principles such as fairness, accountability, and transparency. Second, regulatory requirements: many international regulations emphasise transparency in automated systems. Third, increased user trust: users are more likely to trust AI systems when they understand the reasoning behind predictions. Fourth, improved model validation: explainability enables experts to verify whether model predictions align with domain knowledge. Fifth, bias detection: by analysing feature contributions and explanations, researchers can detect and correct model biases.

1.3 Objectives

The primary objective of this project is to develop an explainable deep learning model that supports ethical and transparent decision support systems. The specific objectives are:

(i)To study existing explainable AI techniques and deep learning models;

(ii) To design a deep learning framework capable of generating interpretable explanations;

(iii) To implement the proposed framework using appropriate tools and datasets;

(iv) To evaluate the performance of the model using accuracy and explanation metrics;

(v) To analyze ethical aspects such as bias detection and fairness in predictions; and

(vi) To assess the usefulness of generated explanations for decision-making.

1.4 Scope of the Project

The scope of this project includes the design, implementation, and evaluation of explainable deep learning models in decision support systems. The project focuses on domains where transparency and ethical considerations are particularly important, such as healthcare and financial decision-making. The main activities include studying state-of-the-art explainability techniques, designing deep learning models with interpretability mechanisms, evaluating model performance using real or publicly available datasets, and analysing ethical aspects such as fairness and bias detection.

1.5 Literature Review

Ribeiro et al. introduced LIME (Local Interpretable Model-Agnostic Explanations), a widely adopted technique that explains model predictions by approximating the behaviour of complex models with simpler interpretable models around a specific

prediction instance. Lundberg and Lee proposed SHAP (SHapley Additive exPlanations), based on cooperative game theory, which assigns importance values to each input feature by evaluating its contribution to the final prediction. SHAP provides a consistent and mathematically grounded approach for understanding feature importance and has become one of the most widely used explainability tools in modern AI practice.

Selvaraju et al. developed Grad-CAM, a gradient-based visualisation technique that highlights areas of input data most strongly influencing model predictions. In image classification tasks, Grad-CAM can identify the specific regions of an image that contribute most to the model's decision. Vaswani et al. introduced the Transformer architecture leveraging self-attention mechanisms, enabling models to focus on the most relevant parts of input data. Doshi-Velez and Kim formalised the science of interpretable machine learning, emphasising the need for rigorous evaluation frameworks for explanation quality beyond simple accuracy metrics. Caruana et al. demonstrated intelligible models in healthcare for predicting pneumonia risk, showing that interpretable AI can achieve performance comparable to black-box models in critical domains. These foundational works collectively establish the theoretical and practical basis for the proposed explainable decision support framework.

2. EXISTING SYSTEM

2.1 Traditional Decision Support Systems

Traditional decision support systems relied mainly on rule-based models and statistical techniques. These systems were designed using predefined rules created by domain experts and were therefore transparent and interpretable. Although these systems were easy to understand, they had several limitations. They were not capable of handling complex data patterns and large datasets. As data volume and complexity increased, traditional systems struggled to maintain accuracy and efficiency.

2.2 Deep Learning Based Decision Support Systems

With advancements in artificial intelligence, deep learning models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been integrated into decision support systems. These models can process large datasets and automatically learn complex relationships within the data. Deep learning models significantly improved prediction accuracy and performance in various domains. For example, they are used for medical diagnosis, fraud detection, credit risk analysis, and autonomous vehicle navigation.

Despite their advantages, deep learning models suffer from a major drawback: the lack of interpretability. The internal operations of these models are difficult to understand because they consist of multiple layers and complex mathematical computations. The existing architecture generally follows a standard workflow where data is collected, preprocessed, and passed to a deep learning model that generates predictions — without any explanation layer for users.

2.3 Explainability Techniques in Deep Learning

Several techniques have been proposed to improve the interpretability of deep learning models. Local Interpretable Model-Agnostic Explanations (LIME) explains model predictions by approximating the behaviour of complex models with simpler interpretable models around a specific prediction instance. SHapley Additive exPlanations (SHAP), based on cooperative game theory, assigns importance values to each input feature by evaluating its contribution to the final prediction, providing a consistent and mathematically grounded approach for understanding feature importance.

Visualisation-based explanation techniques such as gradient-based visualisation and activation mapping highlight the areas of input data that most strongly influence model predictions. Attention mechanisms allow deep learning models to focus on the most relevant parts of input data when generating predictions. By examining attention weights, researchers can identify which input features play the most significant role in the model's decision-making process.

2.4 Limitations of Existing Systems

The existing systems that rely on deep learning models face several critical challenges. The most significant limitation is the lack of transparency — the system provides predictions but does not clearly explain how those predictions were generated. Another limitation is the presence of bias in training data, which may cause the model to produce unfair decisions. User trust is also a major concern, as users may hesitate to rely on outputs they do not understand. Regulatory requirements in many industries now demand transparency in

automated decision-making systems, yet existing black-box systems often fail to meet these standards. Research gaps include the lack of standardised evaluation metrics for explanation quality, scalability challenges, and the need for user-centred explanation design accessible to non-technical domain professionals.

3. PROPOSED SYSTEM

3.1 System Overview

The proposed system aims to develop an Explainable Deep Learning Model for Ethical and Transparent Decision Support Systems. The main objective is to design a system that not only provides accurate predictions but also explains the reasoning behind those predictions. Explainable artificial intelligence techniques are integrated into the system to improve transparency and interpretability. The proposed system also incorporates fairness and ethical considerations to reduce bias and improve reliability.

3.2 Proposed System Architecture

The proposed system architecture integrates explainability techniques within the deep learning model to generate interpretable predictions. The architecture consists of the following major components:

Data Acquisition Module: responsible for collecting datasets from various sources such as healthcare records, financial datasets, or structured tabular data. It performs data collection, validation, and storage for further processing.

Data Preprocessing Module: prepares the raw dataset for model training and prediction. Preprocessing tasks include handling missing data, data normalisation, feature scaling, encoding categorical variables, removing redundant attributes, and splitting datasets into training and testing sets.

Bias Detection Module: analyses the dataset for potential biases before model training, ensuring that the deep learning model does not replicate existing discriminatory patterns present in historical data.

Explainable Deep Learning Model Module: the core computational component. It trains neural network models, learns patterns from historical datasets, and generates predictive models. The trained model is stored for future predictions and analysis.

Explanation Generator (SHAP): the explainability module that provides interpretability to the predictions generated by the deep learning model. It uses SHAP to identify the most important features influencing the prediction, generates feature importance scores, and produces visual explanations understandable to users.

Decision Support Interface: presents results and explanations in graphical formats including prediction outputs, feature importance graphs, and model performance charts, enhancing user understanding of the decision-making process.

3.3 System Workflow

The workflow of the proposed system proceeds as follows. The process begins when the user accesses the decision support interface and uploads a dataset or selects an input sample for analysis. Once the data is provided, it is passed to the data preprocessing module where missing values are handled, categorical variables are encoded, and the dataset is normalised to ensure compatibility with the deep learning model.

After preprocessing, the cleaned data is forwarded to the deep learning model. The model analyses the input data and generates predictions based on patterns learned during training. These predictions represent the system's decision or classification output. To ensure transparency and interpretability, the prediction is then processed by the explainability engine, which uses SHAP to identify the features that have the most influence on the model's decision.

The explanation module generates feature importance values and produces visual representations that help users understand the reasoning behind the model's predictions. Finally, the user reviews the prediction and the corresponding explanation, allowing them to make informed and transparent decisions based on the insights provided by the system.

3.4 Design and Methodology

The system design follows a structured pipeline. The class diagram defines six main classes: User Interface Class, Data Processing Class, Model Training Class, Prediction Class, Explainability Class, and Visualisation Class. The sequence diagram illustrates the interaction between system components beginning with user dataset upload, progressing through the preprocessing module, the deep learning model, and finally the explainability engine. The control flow diagram maps the logical sequence: User Uploads Dataset → Data Preprocessing → Model Training → Prediction Generation → Explainability Analysis (SHAP) → Visualisation of Results → Decision Support Output. The

Entity Relationship diagram defines relationships between User, Dataset, Trained Model, Prediction Results, and Explanation Output entities.

3.5 Implementation

The system was implemented using the Python programming language along with various machine learning and deep learning libraries. The implementation was organised into modular components.

Table 3.1: Software Tools Used

Tool / Library	Purpose
Python	Programming Language
NumPy	Numerical computations
Pandas	Data manipulation and analysis
TensorFlow / Keras	Deep learning model implementation
Scikit-learn	Data preprocessing and evaluation
Matplotlib	Result visualisation
Seaborn	Confusion matrix visualisation
Joblib	Saving preprocessing models
SHAP	Explainability analysis

The deep learning model architecture consists of: Input Layer → Dense Layer (64 neurons, ReLU activation) → Dropout Layer (rate = 0.3) → Dense Layer (32 neurons, ReLU activation) → Output Layer (Sigmoid activation for binary classification). The model is compiled using the Adam optimiser and binary cross-entropy loss function. Early stopping with patience = 10 is applied to prevent overfitting. The trained model is saved as explainable_dl_model.h5.

Hardware requirements: Processor — Intel Core i7 (minimum), Intel Xeon (recommended); RAM — 16 GB (minimum), 32 GB or higher (recommended); GPU — NVIDIA GTX series (minimum), NVIDIA RTX series (recommended); Storage — 512 GB SSD (minimum), 1 TB SSD (recommended). Software requirements: Operating System — Ubuntu / Windows; Deep Learning Framework — TensorFlow / PyTorch; Explainability Libraries — SHAP, LIME; Data Processing — NumPy, Pandas; Visualisation — Matplotlib, Plotly.

4. RESULTS AND DISCUSSION

4.1 Experimental Setup

The results obtained from the implementation of the proposed explainable deep learning model demonstrate the effectiveness of the system in providing accurate predictions along with interpretable explanations. The developed decision support system was tested using a synthetic dataset containing attributes such as age, income, credit score, and loan amount. These features were used to train the deep learning model to predict a binary target variable representing a decision outcome.

The implementation process involved dataset generation, preprocessing, model training, evaluation, and explainability analysis. After the dataset was generated, it was inspected and preprocessed using feature scaling techniques. The dataset was then divided into training and testing sets to ensure that the model could generalise well to unseen data. During the training phase, a feed-forward deep learning neural network was used with multiple dense layers, ReLU activation functions, and dropout layers to reduce overfitting. The model was trained using the Adam optimiser and binary cross-entropy loss function. Early stopping was implemented to prevent excessive training and improve model performance.

4.2 Model Evaluation Results

After training the model, predictions were generated for the testing dataset and evaluated using various performance metrics. The evaluation results were displayed through a confusion matrix, which provides a detailed comparison between actual and predicted values. The confusion matrix output showed that the model was able to correctly classify a large portion of the dataset. The results indicated that the model achieved a high classification accuracy and was capable of distinguishing between the two classes effectively.

The confusion matrix clearly illustrates the number of true positives (256), true negatives (233), false positives (54), and false negatives (57) on the 600-sample test set. The high number of correct classifications demonstrates that the deep learning model is capable of learning meaningful patterns from the input dataset, confirming that the model is suitable for decision support applications where accurate predictions are required.

Table 4.1: Model Evaluation Results on Test Set (600 samples)

Metric	Value	Metric	Value
True Positives	256	True Negatives	233
False Positives	54	False Negatives	57
Accuracy	~81.5%	Precision	~82.6%
Recall	~81.8%	F1-Score	~82.2%

4.3 Explainability Analysis

In addition to prediction accuracy, an important objective of this project was to improve transparency in deep learning models. Traditional deep learning systems often operate as black-box models, making it difficult for users to understand how predictions are generated. To address this issue, the proposed system incorporated an explainability module using permutation feature importance.

The explainability module analyses how each feature influences the prediction results. From the results obtained, it was observed that income and credit score were among the most influential factors affecting the prediction outcome, while age had a relatively smaller impact. The feature importance rankings obtained are presented in Table 3.

Table 4.2: Feature Importance Generated Using Permutation Importance

Feature	Importance Score	Rank
Income	~0.195	1st (Most Influential)
Credit Score	~0.105	2nd
Loan Amount	~0.075	3rd
Age	~0.030	4th (Least Influential)

This explainability analysis provides valuable insights into the model's behaviour and helps users understand the reasoning behind the system's decisions. By visualising feature importance, the system allows decision makers to identify which factors contribute most significantly to predictions. This transparency increases user trust and improves the interpretability of the decision support system.

4.4 Testing Results

The system was validated using both black-box and white-box testing methodologies. Black-box testing evaluates the system based on its inputs and outputs without considering the internal implementation details. Three black-box test cases were executed: dataset generation (BBT01 — Pass), model prediction generation (BBT02 — Pass), and explainability analysis (BBT03 — Pass). White-box testing evaluates the internal structure of the program and ensures that the logic and code paths are functioning correctly. Three white-box test cases were executed: data preprocessing function (WBT01 — Pass), model training function (WBT02 — Pass), and prediction inference function (WBT03 — Pass). All test cases passed successfully, confirming that the implemented modules perform as expected without any major errors.

4.5 Discussion

The experimental results confirm that the proposed system successfully achieves its dual objectives of predictive accuracy and interpretability. The combination of predictive performance and explainability makes the proposed system suitable for ethical and transparent decision support applications. The results highlight the importance of integrating explainable artificial intelligence techniques into deep learning models. By providing clear insights into model behaviour, explainable AI helps bridge the gap between complex machine learning algorithms and human understanding, ultimately contributing to the development of trustworthy and responsible artificial intelligence systems.

5. CONCLUSION

In this project, an Explainable Deep Learning Model for Ethical and Transparent Decision Support Systems was developed to address the challenge of interpretability in modern artificial intelligence systems. Deep learning models are widely used for prediction and decision support tasks due to their high accuracy and ability to learn complex patterns from large datasets. However, these models often operate as black-box systems, making it difficult for users to understand how decisions are generated. This lack of transparency can reduce trust and raise ethical concerns when such systems are applied in critical domains.

The main objective of this project was to design and implement a deep learning based decision support system that not only produces accurate predictions but also provides meaningful explanations for those predictions. A neural network model was implemented using deep learning techniques along with explainable artificial intelligence methods. The system was designed in a modular architecture consisting of data preprocessing, model training, prediction generation, evaluation, and explainability analysis modules. A synthetic dataset containing relevant attributes such as age, income, credit score, and loan amount was generated and used to train the model.

The experimental results confirmed that integrating explainability techniques into deep learning models significantly improves transparency and trust in

automated decision support systems. By providing clear explanations for predictions, the system allows users to better understand model behaviour and evaluate the reliability of its decisions. The proposed system successfully demonstrates how explainable artificial intelligence can be integrated with deep learning to create ethical, transparent, and trustworthy decision support systems.

Although the proposed system demonstrates promising results, several improvements and extensions can be explored in future research. The current implementation uses a synthetic dataset for demonstration purposes. In future work, the system can be applied to real-world datasets from domains such as healthcare, finance, or risk assessment. Additional explainability techniques such as SHAP, LIME, and attention mechanisms can provide deeper insights into the behaviour of deep learning models. Combining multiple explainability approaches could improve the interpretability of predictions. Development of a graphical user interface or web-based dashboard for interacting with the system would make it more accessible for non-technical users. Future research can also focus on improving model performance by exploring more advanced deep learning architectures such as convolutional neural networks and transformer-based models. Ethical considerations such as fairness and bias detection can be further incorporated with tools that detect potential bias and ensure decisions remain fair across different demographic groups.

5.1 Research Gaps

Despite significant advancements in explainable AI, several challenges remain unresolved. One major challenge is balancing model accuracy with interpretability. Highly complex deep learning models often achieve better predictive performance but are considerably more difficult to explain. Another challenge is the lack of standardised evaluation metrics for explanation quality. While prediction accuracy can be measured using well-established statistical metrics, evaluating the usefulness and clarity of explanations remains inherently more subjective and context-dependent.

User-centred explanation design is also an important area for further research. Explanations must be understandable not only to machine learning experts but also to domain professionals and non-technical users who rely on these systems for critical decisions. Scalability presents another significant challenge when applying explanation techniques to large-scale datasets or complex models, as some explanation methods require substantial computational resources, making real-time deployment in production systems difficult. These gaps directly motivate the objectives and design choices of the proposed system described in the following section.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You? Explaining the Predictions of Any Classifier." Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144, 2016.

[2] Lundberg, Scott M., and Su-In Lee. "A Unified Approach to Interpreting Model Predictions." Advances in Neural Information Processing Systems, 30, 2017.

[3] Esteva, Andre, et al. "Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks." Nature, 542, pp. 115–118, 2017.

[4] Doshi-Velez, Finale, and Been Kim. "Towards a Rigorous Science of Interpretable Machine Learning." arXiv Preprint arXiv:1702.08608, 2017.

[5] Vaswani, Ashish, et al. "Attention Is All You Need." Advances in Neural Information Processing Systems, pp. 5998–6008, 2017.

[6] Selvaraju, Ramprasaath R., et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization." IEEE International Conference on Computer Vision (ICCV), 2017.

[7] Sundararajan, Mukund, Ankur Taly, and Qiqi Yan. "Axiomatic Attribution for Deep Networks." Proceedings of the 34th International Conference on Machine Learning, pp. 3319–3328, 2017.

[8] Zhang, Brian Hu, Blake Lemoine, and Margaret Mitchell. "Mitigating Unwanted Biases with Adversarial Learning." Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2018.

[9] Caruana, Rich, et al. "Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-Day Readmission." Proceedings of the 21st ACM SIGKDD, pp. 1721–1730, 2015.

[10] Jain, Sarthak, and Byron C. Wallace. "Attention Is Not Explanation." Proceedings of the North American Chapter of the Association for Computational Linguistics, pp. 3543–3556, 2019.