DEPARTMENT OF BIOMEDICAL ENGINEERING

24AI671-FUNDAMENTALS OF DATA SCIENCE

UNIT IV- DATA ANALYTICS AND MACHINE LEARNING FUNDAMENTALS

4.8. EXPLAINABLE AI (XAI)

Explainable AI (XAI) is a set of methods and techniques that make the **decisions of AI and machine learning models understandable to humans**. It answers the questions: **why did the model give this result, which inputs influenced it, and can we trust it?**

Many accurate models such as deep neural networks, random forests and boosted trees are **black boxes** – we can see the input and output but not the reasoning inside. This is a serious problem when AI is used in healthcare, banking, law and safety-critical engineering, where wrong or unfair decisions can harm people.

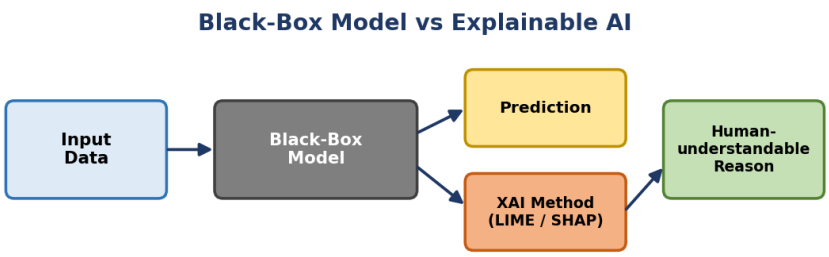


Fig 8.1 XAI explains the output of a black-box model

Why Explainable AI is Needed

- **Trust** – users and engineers accept AI results only if they understand them.
- **Fairness and bias detection** – explanations show if the model uses unfair factors such as gender or caste.

- **Debugging and improvement** – helps developers find errors, wrong features or data leakage.
- **Legal and ethical requirement** – rules like the GDPR ('right to explanation') and AI governance frameworks demand transparency.
- **Safety** – in critical systems, engineers must know why a fault or a danger was predicted.

Interpretability vs Explainability

- **Interpretability** – the model is simple enough to be understood directly (linear regression, decision tree).
- **Explainability** – extra methods are used to explain a complex model after it is trained (LIME, SHAP).

Types of Explanations

Type	Meaning
Global explanation	Explains the overall behaviour of the model – which features are most important in general
Local explanation	Explains one individual prediction – why this particular loan was rejected
Intrinsic (ante-hoc)	Model is transparent by design – decision tree, linear model, rule-based model
Post-hoc	Explanation is created after training for a black-box model – LIME, SHAP, Grad-CAM
Model-specific / Model-agnostic	Works for one type of model only / works for any model

Popular XAI Techniques

LIME – Local Interpretable Model-agnostic Explanations

LIME explains **one prediction at a time**. It slightly changes (perturbs) the input many times, observes how the black-box output changes and then fits

a **simple, interpretable model (such as linear regression) around that single point**. The simple model's weights show which features were most important for that prediction. It works with any model.

SHAP – SHapley Additive exPlanations

SHAP is based on **Shapley values from cooperative game theory**. It treats each feature as a 'player' and calculates **how much each feature contributes to moving the prediction away from the average prediction**. It gives both local and global explanations and has a strong mathematical basis.

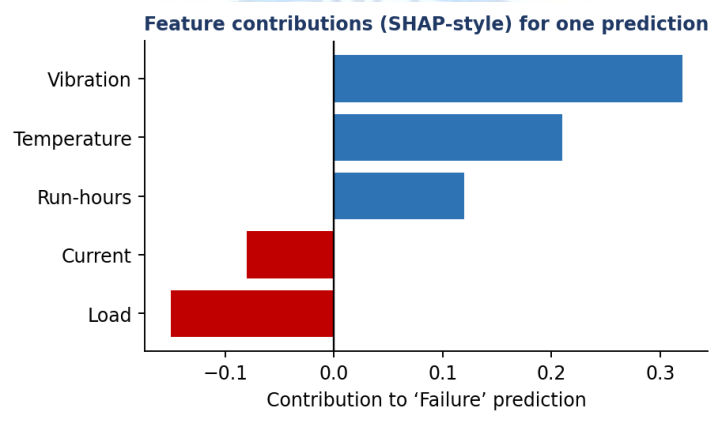


Fig 8.2 SHAP-style explanation of a machine-failure prediction

In the figure, high vibration and temperature push the prediction towards 'failure' (blue bars), whereas high load and current push it away (red bars).

Other Methods

- **Feature importance** – ranks features by how much they affect the output (available in random forests).
- **Partial Dependence Plot (PDP)** – shows how the prediction changes when one feature is changed and others are kept constant.
- **Grad-CAM and saliency maps** – highlight the regions of an image that influenced a CNN's decision.
- **Counterfactual explanations** – show the smallest change in input that would change the decision ('if income was ₹5,000 higher, the loan would be approved').
- **Attention visualisation** – shows which words the transformer focused on.

Advantages and Challenges

Advantages	Challenges
Builds trust and acceptance	Trade-off between accuracy and explainability
Helps find bias and errors	Explanations may be approximate or unstable
Supports legal compliance	Difficult to explain very large deep models
Makes debugging easier	Extra computation time and cost

Applications

- Healthcare – doctors can see which symptoms or image regions led to a diagnosis.
- Banking – reasons for loan approval or rejection and fraud alerts.
- Industry – explaining why a machine was predicted to fail.
- Autonomous vehicles – understanding the car's decisions.
- Recruitment and law – ensuring fair, unbiased decisions.

4.9. AutoML BASICS

Automated Machine Learning (AutoML) is the process of **automating the time-consuming and repetitive steps of building a machine learning model** – from data preparation to model selection and tuning – so that even people with limited ML knowledge can build good models quickly, and experts can save time.

In a normal ML project, a data scientist manually tries many algorithms, creates features and tunes hyper-parameters, which takes a lot of time and skill. AutoML does this search automatically and returns the **best model** for the given data.

AutoML Pipeline

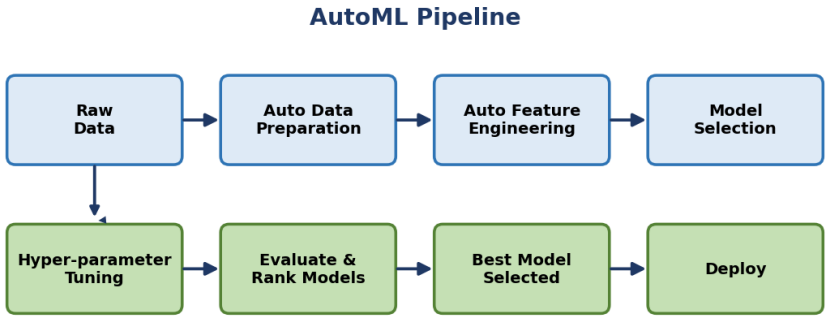


Fig 9.1 Steps automated by AutoML

- **Automated data preparation** – handles missing values, encoding, scaling and outliers.
- **Automated feature engineering** – creates new useful features and selects the best ones.
- **Model selection** – tries many algorithms (linear models, trees, SVM, boosting, neural networks) and compares them.
- **Hyper-parameter optimisation (HPO)** – automatically searches the best settings using grid search, random search, **Bayesian optimisation** or evolutionary methods.
- **Evaluation and ranking** – models are cross-validated and ranked on a leaderboard using the chosen metric.
- **Ensembling and deployment** – the best models may be combined (stacking) and exported for use.

Key Concepts

- **Neural Architecture Search (NAS)** – automatically designs the best structure of a deep neural network.
- **Meta-learning ('learning to learn')** – uses experience from earlier tasks to quickly choose good models for a new task.
- **Hyper-parameter tuning methods:** Grid search tries all combinations; random search tries random combinations (faster); Bayesian optimisation uses past results to choose the next best trial.

Popular AutoML Tools

Tool	Provider / Type	Notes
Auto-sklearn	Open source (Python)	Automates scikit-learn pipelines
TPOT	Open source (Python)	Uses genetic programming to find pipelines
H2O AutoML	Open source / H2O.ai	Fast; produces a leaderboard of models
AutoKeras	Open source	AutoML for deep learning
Google Cloud AutoML / Vertex AI	Cloud service	Easy no-code model building
Azure AutoML, Amazon SageMaker Autopilot	Cloud service	Integrated with cloud platforms

Advantages and Limitations

Advantages	Limitations

Advantages	Limitations
Saves time and effort	High computing cost for large searches
Makes ML easy for non-experts	Less control over the model; may behave like a black box
Often gives good, competitive models	Cannot replace domain knowledge and good data
Reduces human bias in model selection	Risk of overfitting if validation is poor

Applications

- Quick prototyping of predictive models in business and engineering.
- Demand and sales forecasting, churn prediction and fraud detection.
- Predictive maintenance models for industrial machines.
- Image and text classification by non-programmers using no-code platforms.