Mahavir Education Trust's
SHAH & ANCHOR KUTCHHI ENGINEERING COLLEGE
Chembur, Mumbai - 400 088
UG Program in Information Technology

- **Student Name: Purva Mahamunkar**
- **Roll Number: 12/28**
- **PRN : 22UF17661IT090**
- **Department: IT**
- **Batch/Division: BTECH-1**
- **Experiment Title: Explainable AI in Healthcare for Model Interpretation**
- **Date: 18/10/2025**

Experiment No. 8				
Date of Performance:	**18/10/2025**			
Date of Submission:	**25/10/2025**			
Program Execution/ formation / correction/ ethical practices	Timely Submission	Viva	Experiment Total	Sign with Date

Experiment No 8

8.1 Aim:
Explainable AI in Healthcare for Model Interpretation

8.2 Course Outcome:
Develop AI and ML solutions for healthcare applications while collaborating effectively in multidisciplinary teams and adhering to ethical standards and protocols.

8.3 Learning Outcome:
Interpret AI models and visualize decisions for transparency and accountability in healthcare.

8.4 Requirements:
Python, Jupyter Notebook (Google Colab), Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn, SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-agnostic Explanations).

8.5 Related Theory:
Introduction to Explainable AI (XAI) in Healthcare

The rapid adoption of Artificial Intelligence (AI) and Machine Learning (ML), particularly Deep Learning, has revolutionized many industries. However, the most successful models often operate as "black-boxes," meaning their complex internal logic, built from millions of non-linear interactions, is opaque to human understanding. This lack of transparency poses a critical challenge in high-stakes domains like healthcare, where decisions directly impact patient lives and well-being. The consequences of an unexplained erroneous prediction—whether a missed diagnosis or an unnecessary intervention—are severe, leading to legal, ethical, and clinical failures. Explainable AI

(XAI) is an emerging field of research focused on making AI models more comprehensible, reliable, and trustworthy. XAI seeks to bridge the gap between model performance and human understanding by producing explanations that are accessible and meaningful to stakeholders, including clinicians, regulators, patients, and developers. In healthcare, XAI is not merely a technical refinement but an ethical imperative necessary for upholding the principles of beneficence and non-maleficence. Without XAI, AI systems cannot be held accountable, nor can their outputs be validated by expert physicians, fundamentally limiting their integration into clinical workflows.

The Necessity of Transparency, Trust, and Accountability in Healthcare AI

The ethical and practical integration of AI into clinical practice hinges on three pillars: Transparency, Trust, and Accountability.

1. Transparency: This is the ability to understand how an AI model arrived at a specific decision. Clinicians must be able to confirm that a diagnosis recommendation is based on relevant clinical markers (e.g., patient age, lab results, medical history) and not on spurious, non-causal features (e.g., hospital affiliation, insurance type, or room number, as sometimes revealed in predictive models). Transparency facilitates the identification of algorithmic bias and data leakage, which are critical risks in sensitive healthcare data. It moves the technology beyond simply saying "this is the answer" to "this is the answer because..."

2. Trust: Trust is earned when AI consistently provides accurate, validated, and explicable results. A physician will only trust an AI-driven cancer risk prediction if they understand the factors the model considered and agree they align with their clinical expertise. Trust is essential for user adoption; a doctor is highly unlikely to follow a "black box" recommendation that contradicts their professional judgment, thereby defeating the purpose of the AI tool. For patients, trust is built when they can be informed about *why* the AI recommended a specific treatment plan for them.

3. Accountability: Accountability ensures that there is a clear mechanism for determining responsibility when an AI system makes an error. Since the AI itself cannot be held legally or ethically responsible, the burden falls on the developer, the hospital, or the clinician who acted on the advice. XAI provides the necessary audit trail—a detailed, feature-specific breakdown of the decision—that allows human experts to scrutinize the reasoning retrospectively. This is vital for litigation, regulatory compliance (e.g., FDA approval), and continuous quality improvement in clinical pathways.

Model Interpretation Methods and Working Principles

XAI methods are generally categorized into two types based on their scope and application: Global (explaining the entire model behavior) and Local (explaining a single prediction). They are also categorized as Model-Agnostic (works with any ML model) or Model-Specific (only works with certain model types, like tree-based models).

A. Local and Global Interpretation with SHAP (SHapley Additive exPlanations)

Working Principle: SHAP is a unified framework that utilizes concepts from cooperative game theory to explain the output of any machine learning model. The core idea is based on Shapley values, a concept invented by Lloyd Shapley to fairly distribute the payout among a team of players based on their individual contributions to the overall result. In the context of ML, the "players" are the input features, and the "payout" is the prediction. SHAP calculates the average marginal contribution of a feature value across all possible feature combinations (or "coalitions"). This ensures that the sum of the SHAP values for all features equals the difference between the model's prediction for the instance and the average (or base) prediction for the entire dataset. This property, known as local accuracy and consistency, makes SHAP highly mathematically rigorous and fair.

SHAP Workflow:

1. Define Explainer: An appropriate SHAP explainer is initialized (e.g., TreeExplainer for tree models like Random Forest, which is extremely fast and exact; or KernelExplainer for model-agnostic use).

2. Calculate SHAP Values: The explainer computes the contribution of each feature for every instance in the dataset (or a relevant subset, like the test set). For multi-class classification, a set of SHAP values is generated for each class.

3. Local Visualization (Force Plot): This plot visualizes the specific features that push the prediction higher (in red) or lower (in blue) than the average prediction for a single patient. It is arguably the most powerful tool for conveying local, accountable reasoning to a clinician.

4. Global Visualization (Summary Plot / Beeswarm Plot): By aggregating the absolute SHAP values across all patients, this shows the overall feature importance. The plot can also show the direction of influence (e.g., high 'Age' has a high SHAP value, consistently pushing the prediction towards an 'Abnormal' result).

Advantages in Healthcare: SHAP's mathematical rigor and the ability to link local and global insights are paramount. It allows a physician to see not just *that* Age is an important feature, but specifically *how* a patient's Age of 78 contributed a quantified amount (e.g., 20%) to their predicted disease risk. This provides a detailed, quantified justification necessary for high-stakes decisions and is highly effective for both transparency and legal auditability.

B. Local Interpretation with LIME (Local Interpretable Model-agnostic Explanations)

Working Principle: LIME is a simpler, yet highly effective, model-agnostic technique that focuses purely on local fidelity. Its goal is to understand *why* the black-box model made a specific prediction for a given instance by approximating the complex model's behavior around that instance with a simple, interpretable model (like linear regression). LIME is a pragmatic solution when the underlying black-box model (e.g., a complex deep neural network) is too difficult for SHAP to handle efficiently.

LIME Workflow:

1. Instance Selection: A single instance (patient data) is selected for explanation.

2. Perturbation and Prediction: LIME generates numerous slightly modified versions (perturbations) of the original instance by randomly changing the feature values. The black-box model then makes predictions for all these perturbed instances.

3. Local Weighting: LIME assigns weights to these perturbed data points based on their proximity (similarity) to the original instance. Perturbed samples closer to the original sample are given higher weights.

4. Local Model Training: A simple, interpretable model (e.g., a linear model) is trained on the weighted, perturbed data and their corresponding predictions. The simple model's coefficients are then used as the explanation, under the assumption that, locally, the complex model behaves linearly.

5. Visualization: The output displays the top features that contributed most to the prediction, along with their coefficients and the direction of influence.

Advantages in Healthcare: LIME's primary strength is its model agnosticism. It can be universally applied to any black-box model without needing to know the internal mechanics, making it suitable for hospital IT departments dealing with AI models from various external vendors. Its local explanation provides immediate, digestible feedback (e.g., "The model predicted 'Normal' because Blood Pressure <120 was a strong negative factor, and Cholesterol <200 was a weak negative factor"), which is intuitive for clinical staff.

Example Visualizations and Applications in Healthcare

Effective XAI relies heavily on visualizations to translate complex model reasoning into meaningful insights.

1. Feature Importance Plots: These plots (e.g., Random Forest MDI, or SHAP Summary Plot) are used to identify systemic model biases or dependencies. If a plot reveals that a non-clinical feature like "Patient ZIP Code" is highly important, it immediately signals a data quality issue or ethical bias (geographical discrimination) that requires intervention.

2. Local Explanation Plots (LIME/SHAP Force Plot): These are patient-centric tools. A SHAP Force Plot showing exactly which features increased or decreased the risk of a specific diagnosis (e.g., stroke) is directly integrated into the electronic health record (EHR) interface. This empowers the physician to use the AI as a diagnostic assistant rather than an oracle.

3. Confusion Matrix: A fundamental visualization for accountability. It shows the true distribution of model errors (False Positives and False Negatives). For instance, in an infectious disease diagnosis model, a high False Negative rate (missing the disease) is a catastrophic error. The Confusion Matrix quantifies these errors, guiding clinical risk assessment and regulatory review.

Applications:

- Diagnostic Support: Explaining an image recognition model's prediction of diabetic retinopathy by highlighting the specific retinal lesions (pixels) that drove the decision.

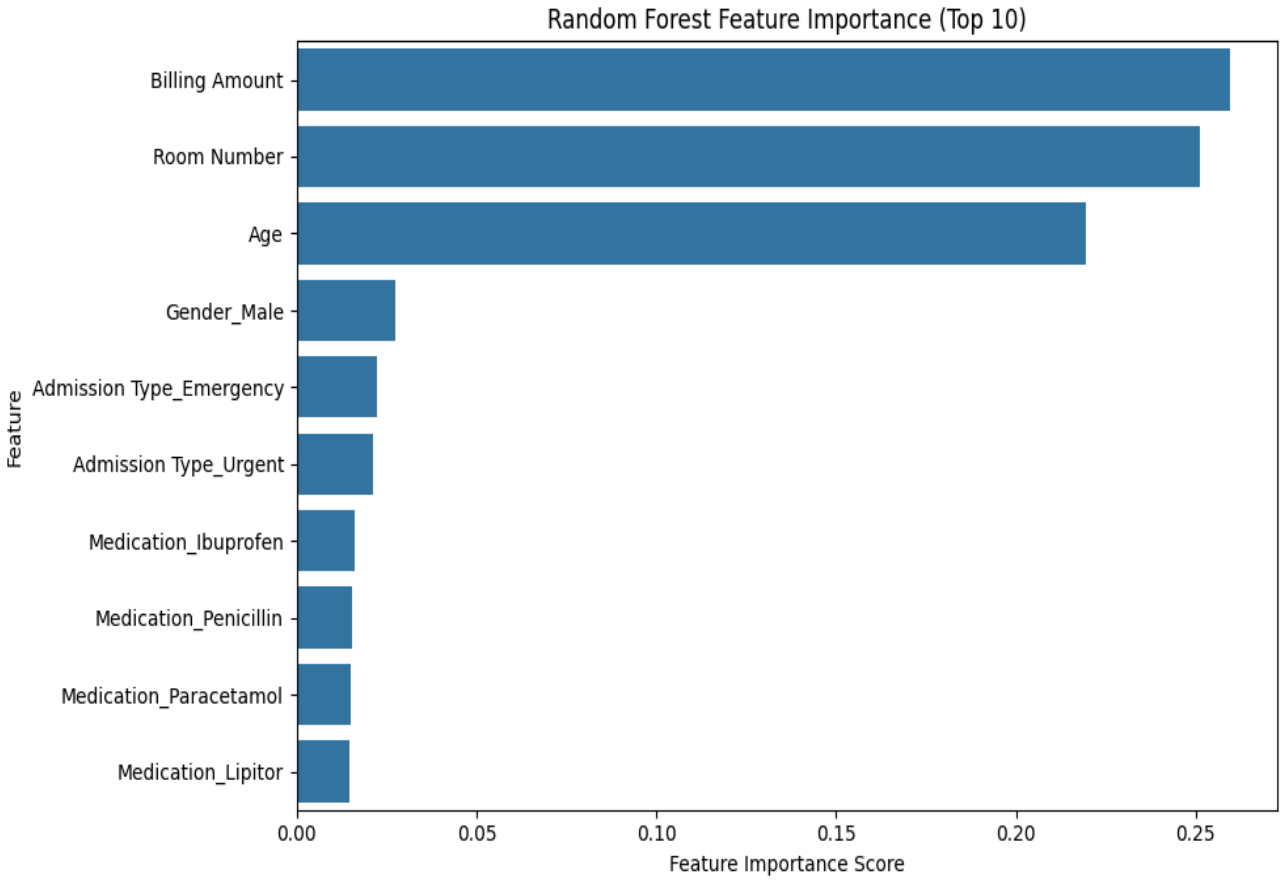
- Risk Prediction: Justifying a patient's predicted risk of hospital readmission by citing contributing features like length of previous stay, medication adherence, and comorbidities, allowing for targeted post-discharge care.

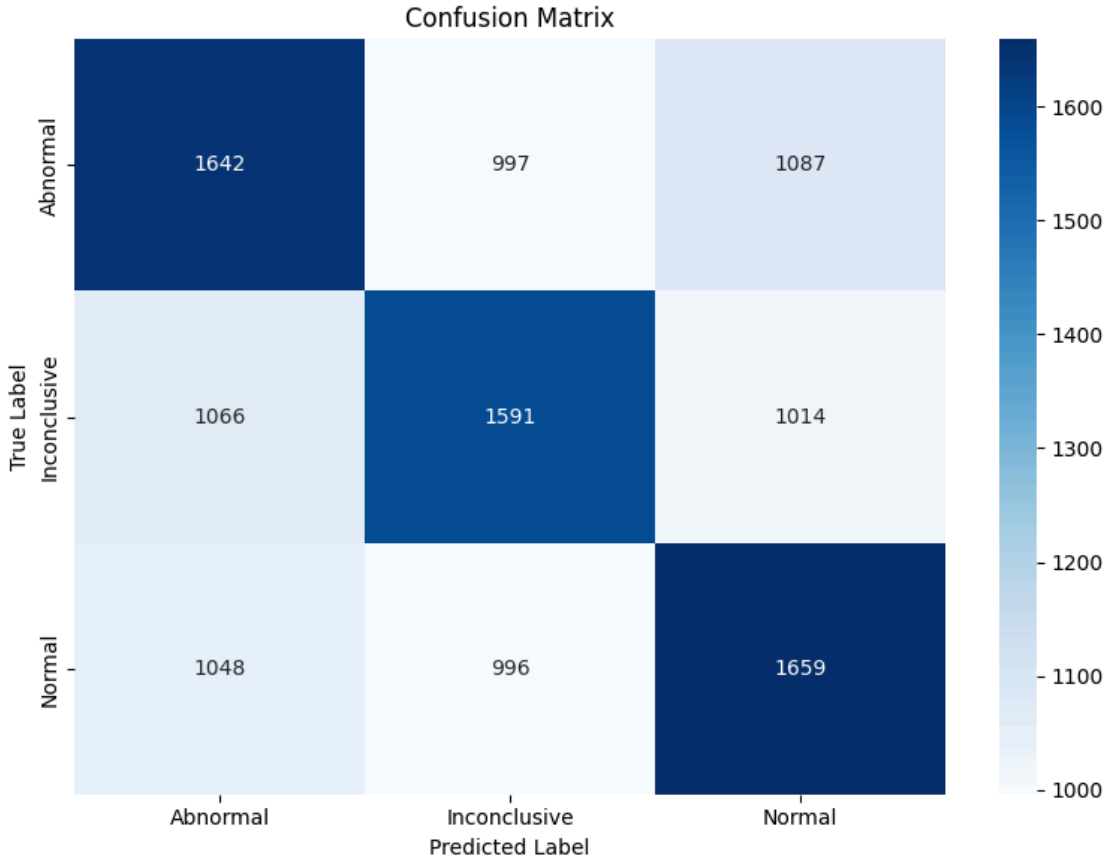
- Ethical Auditing: Using XAI to verify that a model predicting treatment efficacy is not exhibiting demographic bias (e.g., if the model systematically provides less accurate predictions for a specific race or socioeconomic group), ensuring equitable healthcare delivery.

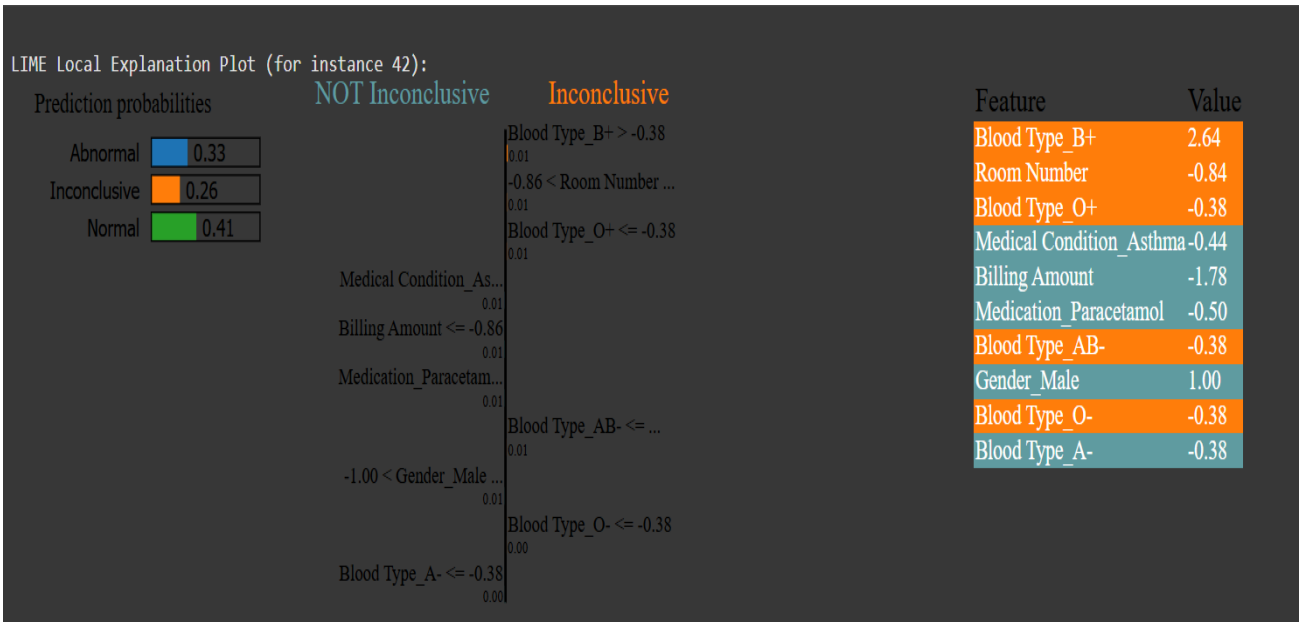
- Accelerating Research: Providing insight into which molecular features or targets an ML model used to predict the efficacy of a new compound, accelerating the drug discovery and validation process.

8.6 Program and Output:

Google Colab Link




Confusion Matrix
Abnormal
Inconclusive
Normal
True Label
1642
997
1087
1066
1591
1014
1048
996
1659
1600
1500
1400
1300
1200
1100
1000
Abnormal
Inconclusive
Normal
Predicted Label


LIME Local Explanation Plot (for instance 42):
Prediction probabilities
Abnormal 0.33
Inconclusive 0.26
Normal 0.41
NOT Inconclusive
Inconclusive
Blood Type_B+ > -0.38
0.01
-0.86 < Room Number ...
0.01
Blood Type_O+ <= -0.38
0.01
Medical Condition_As...
0.01
Billing Amount <= -0.86
0.01
Medication_Paracetam...
0.01
Blood Type_AB- <= ...
0.01
-1.00 < Gender_Male ...
0.01
Blood Type_O- <= -0.38
0.00
Blood Type_A- <= -0.38
0.00
Feature
Value
Blood Type_B+ 2.64
Room Number -0.84
Blood Type_O+ -0.38
Medical Condition_Asthma -0.44
Billing Amount -1.78
Medication_Paracetamol -0.50
Blood Type_AB- -0.38
Gender_Male 1.00
Blood Type_O- -0.38
Blood Type_A- -0.38

8.7 Conclusion:

We have successfully understood that model interpretability is crucial beyond the model's modest 0.4407 accuracy. XAI provided transparency by showing that administrative features like 'Billing Amount' and 'Room Number' were the most influential predictors, which signals potential data bias. Furthermore, the Confusion Matrix delivered accountability by visualizing a key model flaw: high confusion between 'Inconclusive' and 'Abnormal' results.

8.8 Questions:

1. Why is XAI considered an "ethical imperative" in healthcare?

XAI is an ethical imperative because healthcare decisions are high-stakes and directly impact patient safety. Without knowing *why* an AI recommended a diagnosis or treatment, a clinician cannot fulfill their ethical duty of care or maintain autonomy. XAI provides the audit trail necessary to ensure compliance with the principles of beneficence (do good) and non-maleficence (do no harm).

2. What is the fundamental difference between SHAP and LIME?

SHAP is based on game theory (Shapley values) and is mathematically rigorous, guaranteeing that feature contributions sum up to the prediction difference. It offers both local and global explanations. LIME is model-agnostic and provides a local approximation by training a simpler model around the specific data point. It sacrifices mathematical rigor for broader applicability to any black-box model.

3. Why would administrative features like 'Billing Amount' or 'Room Number' become highly important in a healthcare model?

Administrative features often become influential due to data leakage or strong correlation with complexity. For example, a high 'Billing Amount' is often a proxy for a longer hospital stay or a more severe condition, which itself correlates with an 'Abnormal' test result. XAI highlights these proxies, allowing developers to remove them and force the model to rely on true causal clinical factors.

4. How does the Confusion Matrix contribute to accountability in this experiment?

The Confusion Matrix quantifies the types of errors (False Positives/Negatives). In this experiment, it showed exactly how many 'Normal' cases were missed (False Negatives) and incorrectly labeled 'Abnormal.' This provides a concrete, auditable metric on the clinical risk associated with using the model, holding the system accountable for its failure to distinguish between critical and non-critical outcomes.

5. What is the role of transparency in building trust with clinicians regarding AI tools?

Transparency is key to trust. If an AI provides a diagnosis, a clinician must be able to verify that the recommendation is based on their clinical knowledge (e.g., patient's high BMI and old age). If the AI is transparent and its logic is validated by the physician's expertise, the physician is far more likely to trust and adopt the tool, moving from skepticism toward collaborative decision-making.