Part 1: Short Answer Questions (30 points)

1. Problem Definition (6 points)

Hypothetical AI Problem: *Predicting Equipment Failure in Manufacturing Plants*

Objectives:

- Predict potential equipment failures using real-time and historical data.
- Reduce unplanned downtime through predictive maintenance alerts.
- Optimize maintenance schedules to extend equipment life and reduce costs.

Stakeholders:

- Plant Operations Manager – Schedules maintenance efficiently.
- Maintenance Engineers – Perform timely preventive maintenance.

Key Performance Indicator (KPI):

- Accuracy of failure prediction within a specified time

2. Data Collection and Preprocessing (8 points)

Data Sources:

- Sensor Data: Real-time readings (temperature, vibration, pressure).
- Maintenance Logs: Historical records of repairs, failures, and inspections.

Potential Bias in the Data:

Older machines might have more maintenance records than newer ones, leading to over-representation in the training data and biased predictions.

Preprocessing Steps:

- Handling Missing Data: Use interpolation or imputation for missing sensor readings.
- Normalization: Scale sensor readings to a common range for better model performance.
- Label Creation: Assign binary labels (0 = no failure, 1 = failure) based on failure logs.

3. Model Development (8 points)

Model: Random Forest – Suitable for structured data, handles non-linearity, is resistant to overfitting, and provides feature importance insights.

Data Splitting Strategy:

A time-aware split will be used because equipment data (like sensor readings) is time-series in nature, and we must avoid data leakage from the future.

- Training Set (70%): Contains the earliest data, used to train the model on patterns of failure and normal behavior.
- Validation Set (15%): The next slice of data. Used to fine-tune hyperparameters and avoid overfitting.
- Test Set (15%): The most recent data. Used to evaluate how well the model performs on unseen, real-world data.

Hyperparameters to Tune:

- Number of Trees (n_estimators): Affects accuracy and training time.
- Maximum Depth (max_depth): Prevents overfitting by limiting tree complexity.

---

4. Evaluation & Deployment (8 points)

Evaluation Metrics:

1. Precision: Measures how many predicted failures were actual failures (important to avoid false alarms).
2. Recall: Measures how many actual failures were correctly predicted (critical for safety and downtime).

Concept Drift:

A situation where the statistical properties of the target variable change over time (e.g., equipment behavior changes due to aging).

Monitoring Strategy:

Continuously compare model predictions with actual outcomes and retrain the model regularly using new data.

Technical Challenge During Deployment:

Scalability : Processing and predicting from multiple high-frequency sensor streams in real-time requires efficient streaming infrastructure (e.g., Kafka + real-time model inference).

Part 2: Case Study Application

AI-Powered 30-Day Readmission Risk Prediction

1. Problem Scope (5 points)

Problem Statement:

Hospital readmissions within 30 days of discharge are both costly and potentially avoidable, negatively impacting patient health outcomes and hospital performance metrics. Predicting high-risk patients at discharge can enable targeted interventions.

Objective of the AI System:

- Develop a predictive model that assesses each patient's risk of being readmitted within 30 days of discharge.
- Provide clinicians with interpretable risk scores to support discharge planning and post-discharge care.
- Reduce avoidable readmissions, optimize resource utilization, and improve patient outcomes.
- Integrate seamlessly into hospital workflows without disrupting care delivery.

Key Stakeholders:

- Patients: Benefit from proactive care and better recovery.
- Clinicians: Use risk predictions to guide discharge planning and follow-ups.
- Hospital Admins: Monitor performance, reduce penalties, allocate resources effectively.
- Data Scientists & Engineers: Design, develop, and maintain the AI system.

2. Data Strategy (10 points)

Proposed Data Sources:

- Electronic Health Records (EHR): Diagnosis codes, discharge summaries, comorbidities, prior admissions.
- Patient Demographics: Age, gender, zip code, insurance type.
- Lab Results and Vitals: Lab values, blood pressure, pulse,
- Medication Records: Prescribed medications, adherence history.
- Procedures and Treatments: Surgeries, therapies performed during admission.

Ethical Concerns:

1. Patient Privacy and Confidentiality: Use encryption and strict access control.
2. Informed Consent for AI Usage: Implement transparent policies and opt-in consent.

Preprocessing Pipeline:

- Data Cleaning: Handle missing values and duplicates.
- Feature Engineering: Encode variables, aggregate prior admissions, and create time-based features.
- Normalization: Apply Min-Max or Z-score normalization.
- Dataset Splitting: Train (70%), Validation (15%), Test (15%) with stratified sampling.

3. Model Development (10 points)

Model: XGBoost

Why XGBoost?

- Strong performance on imbalanced data
- Non-linear pattern handling
- Built-in feature importance for interpretability

Evaluation:

Hypothetical Confusion Matrix

	Predicted Readmit	Predicted No Readmit
Actual Readmit	80(TP)	20(FN)
Actual No Readmit	30(FP)	870(TN)

Precision = 80 / (80 + 30) = **72.7%**

Recall = 80 / (80 + 20) = **80%**

Interpretation:

- Precision (72.7%): 72.7% of flagged patients were actually readmitted.
- Recall (80%): Identified 80% of actual readmissions.

4. Deployment (10 points)

Integration Strategy:

- Prediction Mode: Batch (daily predictions).
- EHR Integration: Secure API with OAuth 2.0.
- Clinician Dashboard: Embedded module with risk scores and top features.
- Alert System: Optional discharge risk alerts.

Regulatory Compliance :

- Data Encryption: AES-256 at rest, TLS 1.2+ in transit.
- Access Control: RBAC, multi-factor authentication.
- Audit Trails: Monitor all data and model accesses.
- Anonymization: Replace identifiers before model training.

5. Optimization (5 points)

Overfitting Prevention

Method: Cross-Validation

- Use k-Fold Cross-Validation (e.g., k=5) to ensure robustness.
- Avoids overfitting tied to a single data split.
- Improves model generalization and reliability across unseen data.
- L2 Regularization (lambda) in XGBoost to penalize complex models

-Early Stopping: Training halts when validation recall plateaus

Part 3: Critical Thinking (20 points)

Ethics and Bias

Impact of Biased Training Data

- **Misdiagnosis Risks** : Biased models may overlook critical symptoms prevalent in underrepresented populations, leading to incorrect diagnoses.
- **Treatment Disparities** : AI recommendations may favor certain demographics, neglecting the needs of marginalized groups.
- **Exacerbation of Inequities** : Existing healthcare disparities can be worsened as biased AI systems perpetuate unequal treatment access
- **Mitigation Strategy**
- **Diverse Data Collection** : Ensuring that training datasets are representative of the entire patient population can help reduce bias. This includes actively seeking data from underrepresented groups and incorporating social determinants of health into model training.

Trade-offs

The Trade-off Between Interpretability and Accuracy

The trade-off between interpretability and accuracy is a challenge in ML research:

1. **Traditional Models** : Simple, interpretable models like logistic regression or decision trees often struggle with complex datasets, leading to lower accuracy compared to more sophisticated ML approaches
2. **Black-Box Models** : Deep learning models, while highly accurate, are often criticized for their lack of transparency, making them less suitable for clinical applications where interpretability is paramount
3. **Middle Ground** : Recent research has focused on developing models that balance both accuracy and interpretability. For example, techniques like SHAP (Shapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) provide insights into the decision-making process of complex models without significantly compromising their performance.

Practical Implementation Impact

1. **Clinical Validation** : Many models are validated on benchmark datasets but fail to generalize to real-world clinical settings. Ensuring that models perform well across diverse patient populations is essential for practical implementation
2. **Regulatory Hurdles** : The integration of AI models into clinical workflows requires adherence to regulatory standards, which often emphasize interpretability and transparency.
3. **User Acceptance** : Clinicians may be hesitant to adopt models that lack intuitive explanations, even if they are highly accurate. Addressing this requires not only technical advancements but also education and collaboration between data scientists and healthcare professionals.

References

Adeniran, M., Khan, L., & Thomas, R. (2024). *AI governance in healthcare: Regulatory perspectives and frameworks*. Journal of Digital Health Policy, 18(2), 113–129.

Bhatia, P., Okeke, J., & Wong, L. (2024). *Bias in AI medical systems: Causes and consequences*. Health Equity Journal, 9(1), 44–57.

Cross, A., Patel, M., & Ibrahim, N. (2024). *AI and demographic fairness in clinical treatment planning*. Journal of Biomedical Ethics, 11(1), 88–102.

Ennab, A., & Mcheick, H. (2024). *Improving interpretability of deep learning in healthcare*. Computers in Biology and Medicine, 167, 107551.

Luo, Y., Liu, Q., & Chen, Y. (2019). *Interpretable models versus black-box approaches in health prediction*. Artificial Intelligence in Medicine, 97, 1–9.

Shakir, H., Al-Ghamdi, S., & Osei, D. (2024). *Bridging the gap between accuracy and interpretability in clinical AI*. International Journal of Medical Informatics, 179, 105210.

Shamsutdinova, A., Wen, T., & Kang, Y. (2024). *Evaluating machine learning models in real-world clinical environments*. BMC Medical Informatics and Decision Making, 24(1), 67.

Sripathi, R. (2023). *Mitigating algorithmic bias through inclusive dataset design*. AI and Society, 38(4), 925–939.

Tuấn, L. Q. (2024). *Model explainability methods in medical diagnostics: A review*. Expert Systems with Applications, 233, 120378.

Zahraei, S., & Shakeri, A. (2024). *Consequences of training bias in diagnostic models*. Journal of Clinical AI Research, 6(2), 33–49.

Part 4: Reflection & Workflow Diagram (10 points)

WHAT WAS THE MOST CHALLENGING PART OF THE WORKFLOW? WHY?

The most challenging part of the workflow was data preprocessing and bias mitigation. In the hospital readmission case study, ensuring that the data was clean, representative, and ethically sourced was complex due to:

- Missing or inconsistent data in Electronic Health Records (EHRs)
- Sensitive patient information requiring strict privacy handling
- Bias in historical data, such as underrepresentation of certain demographic groups, which could lead to unfair predictions

These challenges required careful feature engineering, imputation strategies, and fairness-aware preprocessing techniques to ensure the model would generalize well and not perpetuate health disparities.

HOW WOULD YOU IMPROVE YOUR APPROACH WITH MORE TIME/RESOURCES?

With more time and resources, I would:

- Expand the dataset by integrating additional sources (e.g., wearable device data, social determinants of health)
- Conduct fairness audits using tools like Aequitas or IBM AI Fairness 360
- Implement federated learning to train models across hospitals without sharing sensitive data
- Engage domain experts (clinicians, ethicists) to validate features and outputs
- Perform longitudinal monitoring to detect concept drift and retrain the model periodically.

References

Brownlee, J. (2018). *How to split a time series dataset into training and test sets* . Machine Learning Mastery. https://machinelearningmastery.com/how-to-split-a-time-series-dataset-into-training-and-test-sets/

Chicco, D. (2021). Training, validation, and test sets: Why splitting matters. *Machine Learning with Applications, 1* (3), 100–106. https://doi.org/10.1016/j.mlwa.2021.100106

IBM Cloud Docs. (2023). *Best practices for training and evaluating machine learning models* . IBM. https://cloud.ibm.com/docs

IBM Research. (n.d.). *AI Fairness 360* . Retrieved from https://aif360.mybluemix.net/

University of Chicago Center for Data Science and Public Policy. (n.d.). *Aequitas: Bias and Fairness Audit Toolkit* . Retrieved from https://www.datasciencepublicpolicy.org/our-work/tools-guides/aequitas/

Zhang, C., & Ma, Y. (2012). *Ensemble machine learning: Methods and applications* . Springer.

Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., ... & Cardoso, M. J. (2020). The future of digital health with federated learning. *NPJ Digital Medicine, 3* (1), 1–7. https://doi.org/10.1038/s41746-020-00323-1

Johnson, A. E. W., Ghassemi, M. M., Nemati, S., Niehaus, K. E., Clifton, D. A., & Clifford, G. D. (2016). Machine learning and decision support in critical care. *Proceedings of the IEEE, 104* (2), 444–466. https://doi.org/10.1109/JPROC.2015.2501978