AURA

**Team Members**

* Ondela Citywayo
* Sinesipho Sibulo
* Segai Mampshika
* Mthunzi Malebadi
* Vhuvhwano Mawela Masalesa

**Project Documentation: Responsible AI Framework for a Medical Diagnosis System**

**1. AI Use Case: AI Medical Diagnosis System**

The chosen AI application is a **Medical Diagnosis System**.

* **Purpose:** The system is designed to assist healthcare professionals by predicting a potential medical diagnosis based on a list of symptoms provided by a patient or doctor. Its goal is to offer a preliminary analysis that can guide further diagnostic testing and consultation.
* **Functionality:** The system operates using a Python script (RealWorldScenario.py) that employs a Naive Bayes classifier. It works by:
  1. **Vectorizing Text:** Converting the input string of symptoms (e.g., "wheezing, rash, pelvic pain") into numerical data.
  2. **Training:** The model is trained on a small, predefined dataset where specific symptom combinations are mapped to 20 different diagnoses.
  3. **Prediction:** When a new set of symptoms is entered, the trained model calculates the most probable diagnosis from its dataset. It also provides a "confidence score," representing the model's certainty in its prediction.
  4. **Reporting:** The output is saved as a text file (diagnosis\_report.txt) and a simple visual report (diagnosis\_image.png).
* **Affected Parties:**
  1. **Patients:** They are the primary subjects whose health and well-being are directly impacted by the accuracy and reliability of the diagnosis.
  2. **Healthcare Professionals (Doctors, Nurses):** They may use this tool to supplement their professional judgment. An inaccurate tool could lead to misdiagnosis, while a reliable one could speed up the diagnostic process.
  3. **Healthcare Institutions:** Hospitals or clinics that deploy this tool are affected, as its performance impacts their quality of care, reputation, and potential liability.

**2. Ethical and Social Risks**

The deployment of this AI system presents several significant risks:

1. **Risk of Critical Misdiagnosis:** The underlying dataset in the script is extremely small and not representative of real-world medical complexity. As seen in the example output where "wheezing, rash, pelvic pain" results in a diagnosis of "Endometriosis" with a very low confidence of 0.15, the model can produce clinically questionable or nonsensical results. Over-reliance on such a tool could lead to incorrect treatment, delayed care for the actual ailment, or severe harm to the patient.
2. **Algorithmic Bias and Lack of Fairness:** The sample dataset lacks demographic information (age, gender, ethnicity, pre-existing conditions), which are crucial for accurate medical diagnosis. The model may inadvertently develop biases, for instance, by predominantly associating certain conditions with the limited examples provided. For example, "Endometriosis," a condition affecting individuals with female reproductive systems, was predicted from a set of general symptoms, highlighting how the model could misapply diagnoses without proper context.
3. **Lack of Transparency and Trust:** While the system provides a "confidence score," it does not explain *why* it reached a specific diagnosis. For a doctor or patient, simply seeing "Endometriosis" with a confidence of 0.15 is not helpful. This "black box" nature erodes trust and makes it impossible for a medical professional to validate the AI's reasoning, turning a potentially helpful tool into a confusing and untrustworthy one.

**3. Applying Principles of Responsible AI**

To be considered responsible, the system must be redesigned with the following principles in mind, drawing from Google's AI Principles and the University of Toronto's Trustworthy AI Framework.

* **Fair & Inclusive:** The current dataset is the biggest point of failure. To ensure fairness, we would need to retrain the system on a large-scale, diverse, and representative medical dataset. This data must be stratified across different ages, genders, ethnicities, and geographies to ensure the diagnostic patterns do not create or reinforce health disparities.
* **Transparent:** The model must be explainable. Instead of just providing a diagnosis and confidence score, an improved system should highlight which specific symptoms most heavily influenced its prediction. This would allow a doctor to see the AI's "reasoning" and evaluate whether it aligns with established medical knowledge.
* **Accountable:** Clear lines of responsibility must be established. The AI should be designated as an **assistive tool** only, with the final diagnostic authority always resting with a qualified human medical professional. A policy must be in place stating that the AI's output cannot be the sole basis for a clinical decision.
* **Privacy-Preserving:** Patient data is highly sensitive. All data, from symptom inputs to diagnoses, must be anonymized and encrypted both in transit and at rest. The system should operate under strict data governance protocols that comply with health privacy regulations (like HIPAA or POPIA), ensuring data is used only for its intended diagnostic purpose.

**4. Design a Mitigation Strategy**

To address the identified risks, the following technical and policy-based actions are proposed:

| **RISK IDENTIFIED** | **MITIGATION STRATEGY** | **TYPE** |
| --- | --- | --- |
| **Critical Misdiagnosis** | **1. Data Enhancement:** Replace the sample data with a verified, peer-reviewed medical dataset.  **2. Confidence Threshold:** Implement a rule where any diagnosis with a confidence score below a high threshold (e.g., 0.85) is automatically flagged for mandatory review by a human doctor and never shown directly to a patient. | Technical/Policy |
| **Algorithmic Bias** | **1. Bias Audits:** Regularly audit the training data and model predictions for performance disparities across demographic groups.  **2. Diverse Data Sourcing:** Actively source training data from a wide range of clinical settings and patient populations. | Technical/Policy |
| **Lack of Transparency** | **1. Implement Explainable AI (XAI):** Integrate methods like LIME (Local Interpretable Model-agnostic Explanations) to show which symptoms contributed most to a diagnosis.  **2. Clear Communication:** The UI must explicitly state that this is a "Preliminary AI Analysis" and not a final diagnosis. | Technical |
| **Privacy Violations** | **1. User Consent Mechanism:** Before a user inputs symptoms, they must be presented with a clear privacy policy and give explicit consent.  **2. Anonymization Protocol:** Implement a technical process to strip all personally identifiable information from the data before it is processed. | Policy/Technical |

**5. AI Ethics Checklist**

This 10-point checklist should be used by the organization before deploying or updating the AI Medical Diagnosis System.

1. **[✓] Data Representativeness:** Has the training dataset been audited to ensure it represents diverse patient populations (age, gender, ethnicity)?
2. **[✓] Clinical Validation:** Has the model's accuracy been validated against established medical benchmarks and peer-reviewed by qualified medical experts?
3. **[✓] Explainability:** Does the system provide a clear explanation for its predictions, not just a result?
4. **[✓] Human Oversight:** Is there a mandatory workflow for a human expert to review low-confidence or high-risk diagnoses?
5. **[✓] Privacy by Design:** Are all data encryption and anonymization protocols in place and verified?
6. **[✓] User Consent:** Is informed consent obtained from users before their data is processed?
7. **[✓] Model Monitoring:** Is there a system in place to continuously monitor the model's performance in the real world and detect performance degradation or new biases?
8. **[✓] Accountability Framework:** Is it explicitly stated in the terms of use that this is an assistive tool and a human professional is the final decision-maker?
9. **[✓] Error Recourse:** Is there a clear and accessible process for patients or doctors to report and appeal a perceived diagnostic error?
10. **[✓] Social Benefit:** Does the system demonstrably improve healthcare outcomes rather than simply creating efficiencies (e.g., does it help diagnose rare diseases faster)?

**6. Visual Framework: Responsible AI Lifecycle**

The following flowchart illustrates how these responsible AI principles are integrated across the entire lifecycle of the Medical Diagnosis System. This visual can be created using a tool called Canva.

