Functional Document

1. **Introduction:**

   The project focuses on enhancing the prediction of diseases associated with miRNA using hybrid machine learning models and explainable AI. miRNAs are small non-coding RNA molecules that regulate gene expression. Alterations in miRNA expression have been linked to diseases like cancer, Alzheimer's, and cardiovascular diseases. Traditional methods to detect miRNA-disease associations are expensive and time-consuming. This enhanced system integrates advanced preprocessing techniques, hybrid modeling, and explainability tools to provide a quicker, cost-effective, and interpretable solution for early disease detection and personalized treatments.

2. **Product Goal:**

   The primary objective of this project is to develop an advanced machine learning system capable of accurately predicting disease associations with miRNAs using a combination of robust models and explainability tools. The system aims to:

   - Improve diagnostic speed and accuracy using hybrid models (e.g., Voting Classifier with Naive Bayes, SVM, and Random Forest, and an enhanced ANN).
   - Offer explainable predictions with SHAP visualizations to provide healthcare professionals actionable insights for decision-making.
   - Utilize large-scale, balanced data from sources like HMDD v3.2, integrated with SMOTE and PCA, to enhance precision and computational efficiency.

3. **Demography (Users, Location):**

   Users:

   - Healthcare professionals (e.g., doctors, clinicians).
   - Medical researchers, geneticists, and bioinformatics experts.
   - Medical laboratories, diagnostic centers, and biotech firms focused on disease diagnosis and personalized medicine.

   Location: The system is globally applicable, with significant use in regions advancing personalized medicine (e.g., USA, Europe, India). It can be integrated into hospitals, research institutions, and biotech companies.

4. **Business Processes:**

   Data Collection and Pre-Processing:

   - Collect miRNA data from HMDD v3.2.
   - Merge multiple datasets based on common identifiers.
   - Handle missing values by removing incomplete rows.
   - Balance the dataset using SMOTE to address class imbalance.

   Feature Engineering:

   - Perform correlation analysis to identify redundant features.
   - Reduce dimensions using PCA to improve computational performance.

   Model Training and Evaluation:

   - Train hybrid models (Voting Classifier combining Naive Bayes, SVM, and Random Forest, and an enhanced ANN with dropout layers).
   - Evaluate models using metrics like accuracy, precision, and recall.

Explainability Integration:

- Use SHAP visualizations to explain feature contributions to predictions.

Implementation and Integration:

- Deploy the best-performing hybrid model.
- Provide real-time predictions and interpretable insights for clinicians.

5. Features

5.1 Feature #1: Predictive Model for Disease Detection

1. Description: The system employs advanced hybrid models to predict diseases based on miRNA expression profiles. Users can upload miRNA datasets to receive disease association predictions and SHAP-based feature importance visualizations.
2. User Story: As a medical researcher, we want to input a patient's miRNA expression data into the system to predict disease risks and understand contributing factors for better treatment recommendations.

5.2 Feature #2: Hybrid Model Support

1. Description: The system combines multiple models (Voting Classifier with Naive Bayes, SVM, and Random Forest, alongside enhanced ANN) for robust and scalable predictions. Users can choose the most suitable model based on data characteristics.
2. User Story: As a clinician, I want to use the hybrid model for accurate predictions on noisy or complex data, ensuring reliable and actionable diagnoses.

5.3 Feature #3: Explainable AI for Real-Time Data Analysis

1. Description: The system integrates SHAP visualizations to provide interpretable predictions in real-time, offering transparency in model decision-making during patient consultations.
2. User Story: As a doctor, I want to input patient data, receive disease predictions, and view feature contributions to explain diagnosis and treatment options to the patient during the same appointment.

6. Authorization Matrix:

- Admin: Full access, including managing datasets, users, and model configurations.
- Medical Researchers: Can upload data, train models, and view predictions along with SHAP visualizations but cannot modify system configurations.
- Clinicians/Healthcare Providers: Can input patient data, view predictions, and access SHAP-based explanations, but cannot alter underlying models or datasets.
- Patients (Optional): Access to personal miRNA test results and disease predictions only through shared reports by healthcare providers

7. Assumptions:

- Data Availability: Large-scale datasets like HMDD v3.2 are available and continuously updated to maintain model accuracy.
- Hardware and Software Infrastructure: High-performance computing resources are available for processing large datasets and running computationally intensive hybrid models.

- User Expertise: Users are familiar with healthcare terminologies, miRNA research, and machine learning system functionalities.
- Legal Compliance: The system complies with legal standards like HIPAA or GDPR to ensure patient data privacy and confidentiality.
- Explainability Tools: SHAP visualizations are used for feature interpretability, assuming users understand their outputs for clinical decision-making.