Marmara University

Engineering Faculty

Aleyna Çakır - 150820023
Hümeyra Eda Devrim - 150820071
Leen I. A. Shaqalaih - 150121921
Mustafa Tolga Akbaba - 150120001
Selin Aydın - 150120061
Zehra Savaş - 150820062
Pablo Buza Mira - 199524008

REPORT AUTOMATION
(Research of Data)

WEEKS 1 & 2

Literature Review of *"A Tree-based RAG-Agent Recommendation System: A Case Study in Medical Test Data"[1]*

Introduction

Medical test recommendation systems help doctors decide which tests to give patients, but traditional methods have limitations. Some systems rely on simple rules like "if a patient has a fever, check for an infection," while others use similarity-based approaches that compare symptoms to past cases. However, these methods often struggle with complex medical situations where symptoms don't always follow a predictable pattern. A new system called HiRMed takes a smarter approach by using Retrieval-Augmented Generation (RAG) and a hierarchical decision-making process. RAG allows the system to not only retrieve relevant medical knowledge from a database but also generate well-reasoned recommendations based on that knowledge. Instead of making one big decision at once, HiRMed follows a step-by-step tree structure, similar to how a doctor logically narrows down possible diagnoses.

Limitations of Current LLM and RAG Approaches in Medical Test Recommendations and Proposed HiRMed Methods as Solutions

1- Medical diagnosis is a step-by-step hierarchical process that starts broad and progressively becomes more specific, similar to how a doctor thinks, where a doctor first looks at the patient's symptoms, narrows down to possible diseases, and finally decides on the most relevant medical tests and medications. The traditional RAG models do not follow this structured approach. Instead, they retrieve documents based on matching the most relevant input symptoms and return the output; thus, they lack logic decision-making and deep reasoning.

Solution: HiRMed solves this issue using a decision tree structure to mimic how doctors think. Each node in the tree performs reasoning by analyzing the retrieved information and making a medical decision before passing it to the next step instead of just retrieving information from past medical tests. Traditional RAG is like a patient searching via Google for “chest pain” and getting multiple related documents without analyzing which one applies to his/her case. On the other hand, HiRMed is like a detective that carefully examines the information step by step before concluding.

2- Medical diagnosis requires both general (e.g., “fever could be caused by an infection”) and specialized medical knowledge (e.g., “in cardiology, a fever could indicate infective endocarditis, while in pediatrics, it might indicate a viral illness"). RAG cannot identify the different cases when making recommendations, leading to erroneous results.
If a system only relies on general medical knowledge, it might miss specialized tests needed for a particular field, and if it only relies on speciality-specific knowledge, it might overlook broader considerations.

Solution: HiRMed separates knowledge into two layers:

1. Root-level knowledge base that covers broad general medical concepts.
2. Department-specific knowledge Base that focuses on specific medical fields like cardiology, endocrinology, etc.

3- A good medical system should remember previous decisions and utilize them to guide future recommendations. Traditional RAG treats each question separately without considering past decisions.

Solution: HiRMed includes Memory-Augmented Reasoning, which keeps track of what medical tests have already been made to prevent recommending the same test multiple times unnecessarily and how the diagnosis changes over time based on new symptoms.

4- A recommendation system also needs to cover all possibilities while also focusing on the most relevant tests. Traditional models struggle to do both at the same time. If a system focuses too much on coverage, it might ask for too many unnecessary tests, wasting time and resources. If a system focuses too much on being specific, it might miss important tests, leading to misdiagnosis.

Solution: HiRMed uses a fine-tuned LLaMa 3.2-3B model to prioritize the most relevant medical tests, such as urgency.

Models Selection

The three key models used in HiRMed's hierarchical medical test recommendation system are

1. OpenAI text-embedding-ada-002 Embedding Model for Retrieval → Converts text into numerical vectors for efficient search.
2. LLM API GPT-O1 for Reasoning → Interprets retrieved information and generates medical recommendations.
3. Weight Model where LLaMA3.2-3B is used for Prioritization → Determines which tests are most important.

Dataset Selection

HiRMed is trained and evaluated using a clinical dataset of 125,000 outpatient visits from various hospital departments. All patient data is anonymized to comply with privacy regulations. The dataset includes:

- Patient symptoms and initial consultations
- Physician-recommended diagnostic tests
- Follow-up tests and treatment outcomes
- Standardized clinical guidelines

Findings and Results

- HiRMed achieved a 92.3% coverage rate, ensuring a broader range of relevant tests are recommended. This was significantly higher than the 84.7% achieved by Flat-RAG, a single-layer RAG system, and 72.8% from Traditional Vector Similarity (TVS) approaches.
- The system attained 88.7% accuracy, surpassing Flat-RAG's 82.4% and TVS's 71.5%, thus making more precise recommendations.
- HiRMed reduced the omission of critical tests to 2.1%, compared to 5.8% in Flat-RAG and 10.6% in TVS, showing that it can capture essential and critical diagnostic tests.
- A panel of doctors rated HiRMed's recommendations 4.3/5 on clinical relevance, indicating strong alignment with expert decisions. Flat-RAG received 3.7/5, while TVS scored 3.2/5.

Literature Review of ''CPLLM: CLINICAL PREDICTION WITH LARGE LANGUAGE MODELS''[2]

INTRODUCTION

Clinical Prediction

This study aimed to facilitate care by making predictions in the next steps of the patients. The system that can make predictions about clinical disease and re-hospitalization was created with the LLM (kind of AI model) model. It was aimed to diagnose the patient who came to the hospital again with a target disease by using the past diagnosis records. The prediction was strengthened by comparing the analysis results with different sources such as RETAIN, Med-BERT and EHR.

Based on the patient's medical history, procedures, diagnoses and medications, it aims to predict whether the patient will be re-admitted to the hospital in the future.

Data Processing

The database and electronic data obtained from the hospital were combined in the program. The prediction task for the data set revolves around determining whether a patient will be diagnosed with a certain disease at their next visit. The system includes medications, procedures, diagnosis codes and health data that can be used for disease diagnosis. The information uploaded for diseases and medications was converted into software codes.

MIMIC-IV and eICU-CRD data sets were used for three diagnosis predictions, including Chronic Kidney Disease, Acute and Unspecified Renal Failure and Adult Respiratory Failure. For each prediction task, patients with certain disease ICD codes were assigned a positive label and their diagnosis history covered all recorded diagnosis codes until the specific code showed a result.

Diagnosis Prediction Result

CPLLM-Llama2 shows superior performance compared to CPLLM-BioMedLM in this specific task. Logistic Regression outperforms RETAIN in both PR-AUC (35.05%) and ROC-AUC (74.664%), but outperforms Med-BERT in PR-AUC. For acute and unspecified renal failure, CPLLM-Llama2 achieved the highest metrics with a PR-AUC score of 45.442% and a ROC-AUC score of 78.504%. This means a significant improvement of 4.22% in PR-AUC compared to the leading baseline model RETAIN in this task. CPLLM-Llama2 shows superior performance compared to CPLLM-BioMedLM. It is noted that CPLLM-Llama2 outperformed CPLLM-BioMedLM and therefore the remainder of the analysis will be based on CPLLM-Llama2. Evaluations of data run on different models.

Conclusion

In this paper, we present CPLLM, a novel method for clinical disease prediction and patient hospital readmission prediction based on patients' clinical history. CPLLM has potential for practical application. By exceeding the state-of-the-art in clinical task prediction, our method provides more accurate and robust disease prediction as well as patient hospital readmission prediction. CPLLM showed superior performance on both datasets (MIMIC-IV and eICU-CRD). It includes ICD9 and ICD10 diagnoses, procedures, and medications.

Literature Review of “ Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda”[3]

Introduction

Artificial intelligence (AI) is revolutionizing healthcare by improving disease diagnosis, patient management, and predictive analytics. AI techniques such as machine learning (ML) and deep learning (DL) enhance accuracy, speed, and accessibility, particularly in resource-limited settings. AI-based models are reshaping medical research and diagnostics by leveraging medical imaging, genomics, and clinical data. Various ML and DL techniques are being applied across multiple medical fields, improving diagnostic efficiency and patient outcomes.

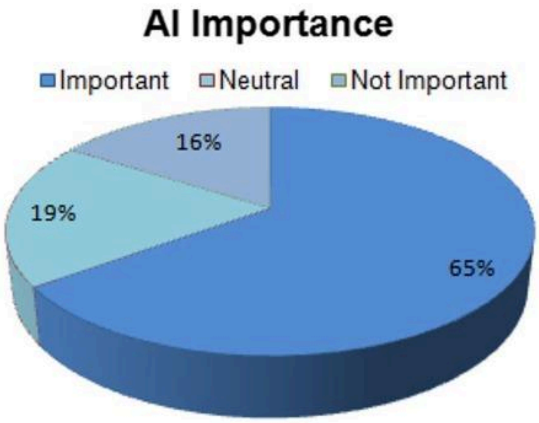


Figure 1: Importance of artificial intelligence in healthcare

Limitations of Traditional Diagnostic Approaches and AI as a Solution

1. Complex Medical Conditions: Traditional diagnostic methods rely heavily on manual analysis and physician expertise, leading to potential misdiagnoses in complex cases. AI-driven models improve accuracy by identifying patterns in large datasets.
2. Data Integration Issues: Traditional methods struggle with integrating diverse medical data sources, such as MRI scans and lab reports. AI enables seamless data processing and real-time insights.
3. Delayed Diagnosis: Early disease detection is often challenging. AI facilitates early diagnosis by analyzing patient histories and predicting risk factors using deep learning models.
4. Inconsistent Results: Physician diagnoses may vary based on experience. AI provides standardized, reproducible results, reducing variability in diagnostic decisions.

AI-Based Diagnostic Framework A structured AI framework for disease diagnosis consists of:

- Data Preprocessing: Cleaning, normalizing, and structuring medical data.
- Feature Extraction: Identifying key indicators for disease classification.
- Machine Learning Models: Implementing classifiers such as Support Vector Machines (SVM), Random Forest, and Neural Networks.
- Evaluation Metrics: Measuring accuracy, sensitivity, specificity, and F1-score.

Challenges in AI-Based Disease Diagnosis

1. Data Privacy & Security: Ensuring compliance with medical regulations and safeguarding patient data.
2. Bias in AI Models: Training models with diverse datasets to prevent biases in disease prediction.
3. Interpretability: Making AI decision-making transparent and explainable for clinical use.
4. Integration with Healthcare Systems: Developing AI models that seamlessly integrate into existing medical workflows.

Future Research Directions

- Personalized Medicine: AI-driven treatment plans based on genetic and lifestyle factors.
- Improved Clinical Decision Support: AI-assisted recommendations for physicians.
- Enhanced AI-Driven Imaging: Refining deep learning models for higher diagnostic accuracy.
- Federated Learning: Secure AI training across multiple medical institutions.

Tuberculosis and AI Applications

AI has been widely used for tuberculosis (TB) detection through imaging techniques like chest X-rays and computed tomography (CT) scans. Convolutional Neural Networks (CNNs) and other deep learning models analyze medical images to identify TB-related abnormalities. Studies have shown that AI models can achieve high sensitivity (up to 97.13%) and specificity (above 80%), making them valuable tools for TB screening in both high- and low-income regions. Raman Spectroscopy combined with machine learning is also being explored to detect TB biomarkers in blood samples.

AI in Other Disease Diagnoses

Beyond TB, AI is extensively utilized in diagnosing various conditions:

- Alzheimer's Disease: AI models analyze neuroimaging data to predict disease progression with up to 96% accuracy. Machine learning methods help detect cognitive decline early, improving intervention strategies.
- Cancer Detection: AI-driven models process histopathology images and genetic data to detect cancer types, assess malignancy levels, and suggest treatment plans. Deep learning achieves over 90% accuracy in detecting lung and breast cancer.
- Heart Disease & Stroke: AI-powered algorithms analyze ECG signals, echocardiograms, and patient health records to identify cardiovascular risks and predict stroke occurrences with high specificity.
- Diabetes & Chronic Illnesses: AI helps monitor glucose levels, predict complications, and personalize treatment plans using data from wearable devices and electronic health records.

Challenges & Future Prospects

While AI has significantly improved disease detection, challenges such as data privacy, model bias, and interpretability remain. Future advancements aim to refine AI models for better generalizability and integration into clinical workflows, ultimately enhancing global healthcare accessibility.

WEEK 3

Literature Review of "Government-Backed Blood Test Analysis for Disease Prediction in Peru" [4]

Introduction

The **Peruvian government** has developed a public dataset containing blood test results linked to various diseases, including diabetes, anemia, obesity, hypertension, and kidney diseases. This initiative provides an open-source platform for researchers to analyze large-scale health data and develop AI-driven disease prediction models.

This study explores the dataset's structure, scope, and potential applications in machine learning for predicting diseases based on blood biomarkers.

Dataset Overview

- **Source:** Peruvian Ministry of Health.
- **Contents:**
  - Patient demographics (age, sex, region).
  - Blood test results (glucose, hemoglobin, creatinine, cholesterol, etc.).
  - Medical diagnoses (linked to disease codes, e.g., diabetes type 2, CKD, hypertension).
  - Metadata and documentation (detailed explanations of column values).
- **Format:** Excel files categorized by disease type, allowing for specialized machine learning applications.

Findings & AI Applications

- **Early detection models**: Using supervised learning, algorithms trained on this dataset can predict disease onset with high accuracy.
- **Public health insights**: Large-scale analysis of blood test patterns provides epidemiological insights into regional disease prevalence.
- **Integration with AI frameworks**: The dataset enables testing of advanced models, including deep learning, decision trees, and XGBoost, for personalized health assessments.

Conclusion

The Peruvian government dataset serves as a valuable resource for researchers aiming to automate disease prediction via blood test data and AI models. Future research should focus on enhancing dataset accessibility, improving feature engineering, and integrating findings into clinical practice.

Literature Review of "Prediction of Coronary Heart Disease Using Routine Blood Tests" [5]

Introduction

This study investigates the potential of routine blood test results to predict the risk of coronary heart disease (CHD). The authors developed a two-layer Gradient Boosting Decision Tree (GBDT) model to classify individuals' health status, distinguishing between healthy individuals, CHD patients, and those with other diseases. The study emphasizes the value of readily available blood test data in assessing CHD risk, offering a non-invasive and cost-effective approach to early detection. Additionally, it compares the model's performance with existing risk prediction methods, such as the Framingham Risk Score and ACC/AHA13, highlighting its advantages in accessibility and affordability.

Data Processing

The researchers utilized a hospital-based cohort comprising 5,060 CHD patients (2,365 men and 2,695 women) aged between 1 and 97 years, with medical records spanning from 2009 to 2017. Additionally, the dataset included 5,051 health check-ups and 5,075 cases of other diseases, originally totaling 16,860 patients. To ensure balanced training, the dataset was adjusted to 15,000 routine blood test results. The GBDT model was trained on this dataset to classify individuals into health status categories accurately.

Findings and Results

The GBDT model demonstrated high sensitivity in detecting CHD. It accurately identified 86% of individuals with the disease, while achieving a 93% sensitivity rate in distinguishing CHD from other diseases in the second classification layer. Key findings indicate a strong correlation between certain blood test markers and CHD risk. The most relevant biomarkers in the first classification layer (healthy vs. disease) were RBC (Red Blood Cell count), HCT (Hematocrit), and LY% (Lymphocyte percentage). In the second layer (CHD vs. other diseases), the most important biomarkers were PDW (Platelet Distribution Width), RDW (Red Cell Distribution Width), and MPV (Mean Platelet Volume).

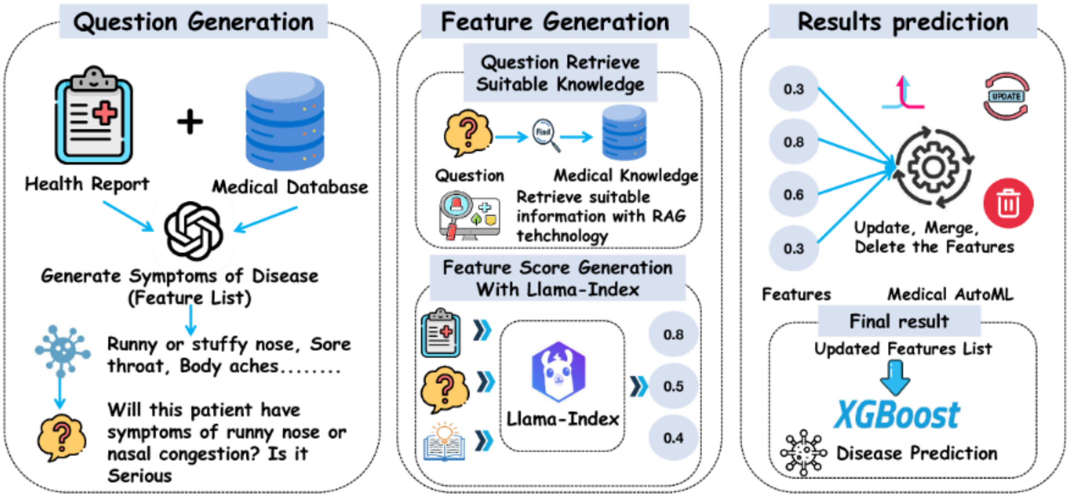
Conclusion

This study underscores the potential of machine learning algorithms, particularly the GBDT model, in utilizing routine blood test data for early prediction of coronary heart disease. By leveraging readily available and non-invasive blood test results, healthcare professionals can enhance early detection strategies, potentially improving patient outcomes through timely interventions. The findings suggest that routine blood test data, often underutilized in CHD screening, could serve as a valuable tool for risk assessment, offering a low-cost and accessible alternative to traditional prediction models.

Literature Review of “Health-LLM: Personalized Retrieval-Augmented Disease Prediction System” [9]

Introduction

Traditional healthcare systems often rely on static data and standardized protocols, which may not fully accommodate individual patient needs. To address this limitation, the authors propose Health-LLM, a system that integrates patient health reports into LLMs[8] to provide detailed task information. This approach allows for the extraction of personalized health characteristics, which are then adjusted using professional medical expertise to improve disease prediction accuracy.



Question Generation
Health Report
Medical Database
Generate Symptoms of Disease (Feature List)
Runny or stuffy nose, Sore throat, Body aches........
Will this patient have symptoms of runny nose or nasal congestion? Is it Serious
Feature Generation
Question Retrieve Suitable Knowledge
Question
Medical Knowledge
Retrieve suitable information with RAG tehchnology
Feature Score Generation With Llama-Index
Llama-Index
0.8
0.5
0.4
Results prediction
0.3
0.8
0.6
0.3
UPDATE
Update, Merge, Delete the Features
Features
Medical AutoML
Final result
Updated Features List
XGBoost
Disease Prediction

Figure 2: The whole framework of Health-LLM [6]

Methodology

Health-LLM's methodology comprises several key components:

1. **In-Context Learning for Symptom Feature Generation**: Utilizes advanced LLMs to systematically extract symptom features from a range of diseases through in-context learning, enabling the model to generate symptom descriptors efficiently.
2. **Retrieval-Augmented Generation (RAG)**: Integrates a supplementary knowledge base using RAG mechanisms to enhance the precision of the generative process by retrieving relevant medical information that aligns with the symptoms mentioned.

3. **Feature Scoring with Llama Index**: Employs the Llama Index [37] framework to assign different weights and scores to health features based on professional medical information, facilitating precise scoring of medical knowledge.
4. **Automated Feature Engineering and Disease Prediction**: Incorporates automated feature engineering techniques to iteratively optimize feature extraction, followed by training an XGBoost classification model [38] for early disease prediction and personalized health recommendations.

Conclusion

The experimental results indicate that Health-LLM surpasses traditional methods in disease prediction accuracy. Specifically, the system achieved an accuracy of 0.833 and an F1 score of 0.762, outperforming models like GPT-4 combined with information retrieval, which had an accuracy of 0.68 and an F1 score of 0.71. These findings suggest that Health-LLM has the potential to revolutionize personalized health management by providing early predictions of potential diseases and customized health advice, thereby enhancing the application of large language models in healthcare.

Literature Review of “Rare disease diagnosis using knowledge guided retrieval augmentation for ChatGPT” [7]

Introduction

Rare diseases, while individually uncommon, collectively impact nearly 400 million people worldwide [11]. The diagnostic process for rare diseases is highly challenging, often taking an average of four to five years for an accurate diagnosis. Many patients remain misdiagnosed or undiagnosed due to the complexity and variability of symptoms, the lack of clinician experience, and inequities in access to specialized diagnostic services.

Machine learning technologies, particularly large language models (LLMs) like ChatGPT, offer potential solutions to improve the efficiency and accuracy of rare disease diagnoses. However, standard LLMs often suffer from "hallucinations"[10]—producing factually incorrect but linguistically coherent outputs. To address this issue, the study introduces *RareDxGPT*, an enhanced version of ChatGPT 3.5, which integrates domain-specific knowledge through Retrieval Augmented Generation (RAG). RareDxGPT retrieves the three most relevant documents from the *RareDis Corpus* and supplies them to ChatGPT to enhance its diagnostic reasoning.

Methodology

The methodology focuses on developing *RareDxGPT* by augmenting ChatGPT 3.5 with the RareDis Corpus [12], a knowledge base containing detailed information about 717 rare diseases, including symptoms, causes, and genetic factors. The process includes:

1. **Retrieval-Augmented Generation (RAG):** Documents in the RareDis Corpus are vectorized and stored in the FAISS (Facebook AI Similarity Search) database [9], allowing efficient semantic retrieval of the most relevant documents when a query is entered.

2. **Semantic Similarity Search:** Queries are matched against the knowledge base using LangChain’s similarity scoring, which utilizes cosine similarity for document retrieval.

3. **Prompt Engineering:** Three different prompt types were tested to optimize diagnostic accuracy:

    - **Basic Prompt:** Providing the patient's symptoms and requesting a diagnosis.
    - **Prompt + Explanation:** Requesting a diagnosis with reasoning.
    - **Prompt + Role Play:** Asking ChatGPT to act as a rare disease specialist to enhance engagement.

4. **Evaluation Metrics:** The performance of RareDxGPT and ChatGPT 3.5 was tested using 30 disease case reports extracted from PubMed. Diagnoses were compared to verified clinical diagnoses, and accuracy was measured. The *Global Quality Scale (GQS)*, a Likert-based assessment, was used to evaluate response quality.

Conclusion

The study finds that *RareDxGPT* outperforms standard ChatGPT 3.5 in diagnostic accuracy for rare diseases. Key results include:

- Using a basic prompt, RareDxGPT achieved a 40% accuracy rate, slightly outperforming ChatGPT 3.5 (37%).
- With the "Prompt + Explanation" method, RareDxGPT reached 43% accuracy, while ChatGPT 3.5 dropped to 23%.
- The "Prompt + Role Play" approach yielded 40% accuracy for RareDxGPT and 23% for ChatGPT 3.5.
- While RareDxGPT provided better diagnostic accuracy, ChatGPT 3.5 produced more natural and well-structured responses according to the GQS ratings.

The findings suggest that integrating domain-specific knowledge via retrieval augmentation can improve ChatGPT's performance in rare disease diagnosis. However, challenges remain, including incomplete knowledge coverage in the RareDis Corpus, limitations in diagnosing conditions with less distinctive symptoms, and ChatGPT's tendency to justify incorrect diagnoses through hallucinations. Future work should focus on expanding the knowledge base, incorporating genomic data, and improving image-based diagnostic capabilities.

WEEK 4

Automated Blood Test Reporting and Recommendation System

1. Introduction

In the field of healthcare, timely and accurate analysis of blood test results is crucial for effective patient management. However, manual interpretation of test results can be time-consuming and prone to human error. To address these challenges, we propose an AI-powered report automation system that not only analyzes blood test results but also provides personalized recommendations for medication and dietary adjustments. [13]

2. System Overview

Our system utilizes machine learning algorithms to interpret blood test parameters and generate automated reports. The key functionalities include:

- Data Processing: The system extracts relevant information from blood test results.
- AI-Powered Analysis: Machine learning models identify deviations from normal ranges and assess potential health risks.
- Personalized Recommendations: Based on the analysis, the system suggests appropriate medications, supplements, or dietary changes.
- Automated Report Generation: A structured and easy-to-understand report is generated for both patients and healthcare providers.

3. Example Use Cases

- Vitamin B12 Deficiency: If a patient's B12 levels are below the optimal range, the system suggests vitamin B12 supplements or B12-rich foods like fish, eggs, and dairy.
- High Cholesterol Levels: In case of elevated cholesterol, the AI recommends lifestyle changes, such as reducing saturated fats and increasing fiber intake.
- Iron Deficiency Anemia: If iron levels are low, the system suggests iron supplements and iron-rich foods like spinach and red meat.

4. Benefits of the System

- Improved Efficiency: Reduces the time needed for manual interpretation and documentation.

- Enhanced Accuracy: Minimizes human errors in diagnosis and recommendations.
- Personalized Healthcare: Tailors recommendations to individual patients based on their unique test results.
- Better Patient Engagement: Provides clear, actionable insights that help patients manage their health effectively.[14]

Conclusion

The implementation of AI-driven automation in blood test analysis has the potential to revolutionize patient care by making diagnosis and treatment recommendations more efficient and accurate. By integrating this technology, healthcare providers can optimize their workflow, improve patient outcomes, and enhance overall healthcare accessibility.

OBSERVATION

The reports to be interpreted by artificial intelligence include blood test samples taken from patients. It is based on the examination of values such as vitamins (such as B vitamins, vitamin D), iron amount, and glucose amount in the blood. Artificial intelligence analyzes will be created based on the test results given by patients in their daily routine and disease states.

The data sets to be used in this study will be the sample results taken from patients. Patient information from hospitals and clinics is protected by the data protection law of Ministry of Health under the law of protection of personal health data with law.[15] Patient information is prohibited from being shared with third parties within the scope of the personal data protection law. The sharing of health data with third parties is protected by decision.[16] This data set creates difficulties in access. An application must be made to the ethics committee in order to share the analysis data of patients, and it can take as long as three months for the results to come from the ethics committee.[17] Access to this kind of data, which contains health data, is quite difficult.

WEEK 5: Dataset Gathering

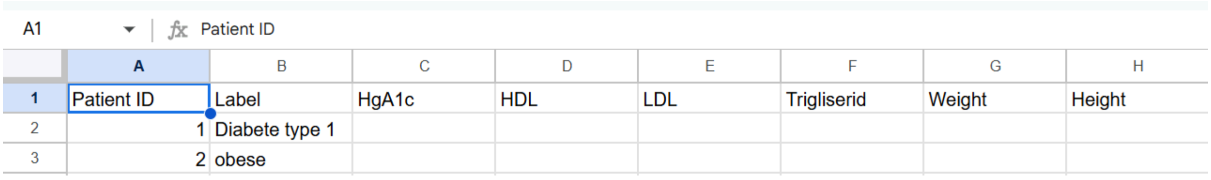
Local Dataset [25]

Our aim for the previous week was to gather blood test reports from local Turkish hospitals to run our project of report automation; however, we were not given access due to the credentials of such data. We obtained a dataset from Üsküdar Diabetes and Obesity Center of Fatih Sultan Mehmet Education and Research Hospital. We reframed our project's goal by analyzing the blood test reports to predict whether the patient is more prone to diabetes.

The dataset contains about 200 patient instances, as well as 8 attributes:

- **Patient ID** – A unique identifier assigned to each patient.
- **Label** – The classification of the patient's condition (e.g., **Diabetes Type I, Type II, Obese**).
- **HbA1c Test** – Also known as **glycated hemoglobin** or **A1c**, this test measures how well a person's blood sugar has been managed over time. It is widely used for diagnosing diabetes and monitoring long-term glucose control **[24].**
- **HDL (High-Density Lipoprotein)** – Often referred to as "good cholesterol," as higher levels reduce cardiovascular risks **[26]**.
- **LDL (Low-Density Lipoprotein)** – Commonly known as "bad cholesterol," as excessive levels can lead to arterial blockages and increase diabetes risk **[27]**.
- **Triglycerides** – A type of fat found in the blood; high levels are linked to metabolic disorders, including diabetes **[28]**.
- **Weight and Height** – Used to calculate the **Body Mass Index (BMI)**, a crucial indicator for obesity, which is a major risk factor for diabetes **[29]**.

Since the dataset is relatively small, it presents limitations in building a robust and well-trained predictive model. To address this, we plan to **merge two additional datasets** with our real dataset, allowing for better generalization and improved accuracy in diabetes prediction. Combining datasets from different sources has been shown to enhance model performance in medical machine learning applications **[30]**



	A	B	C	D	E	F	G	H
1	Patient ID	Label	HgA1c	HDL	LDL	Trigliserid	Weight	Height
2	1	Diabete type 1						
3	2	obese						

Figure 3: Sample data of the local dataset (incomplete). *Full dataset will be received by Saturday.*

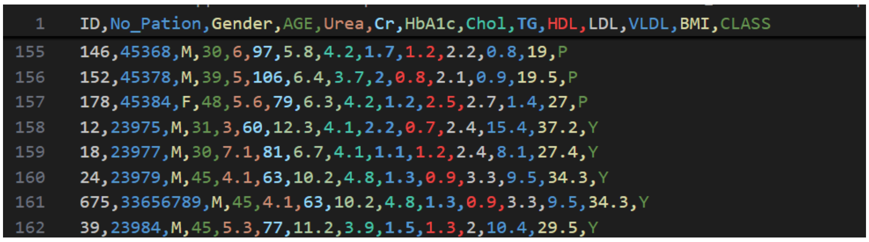
Diabetes Dataset [23]

This dataset comprises medical information and laboratory analyses collected from the Iraqi population in 2020. The data were sourced from the Medical City Hospital laboratory and the Specialized Center for Endocrinology and Diabetes at Al-Kindy Teaching Hospital. Patient files were reviewed, and relevant data were extracted and entered into a database to construct this comprehensive diabetes dataset. It consists of 1000 instances and 14 attributes:

- **ID**: A unique identifier assigned to each patient.
- **No_Pation**: The patient's registry number
- **Gender**: **M (Male)** or **F (Female).**
- **AGE**
- **Urea**: The blood urea nitrogen level, used to assess kidney function and hydration status. [31]
- **Cr (Creatinine)**: A marker of kidney function; high levels indicate impaired renal function. [31]
- **HbA1c (Glycated Hemoglobin)**: Measures average blood sugar levels over the past 2-3 months; a key indicator for diabetes diagnosis. [24]
- **Chol (Total Cholesterol)**: The total amount of cholesterol in the blood, including HDL, LDL, and VLDL. [32]
- **TG (Triglycerides)**: A type of fat found in the blood; high levels are linked to cardiovascular diseases.[28]
- **HDL (High-Density Lipoprotein)**: The "good" cholesterol that helps remove excess cholesterol from the bloodstream. [26]
- **LDL (Low-Density Lipoprotein)**: The "bad" cholesterol that can accumulate in arteries and increase heart disease risk. [27]
- **VLDL (Very Low-Density Lipoprotein)**: A type of lipoprotein that carries triglycerides in the blood; high levels are associated with increased cardiovascular risk. [33]
- **BMI (Body Mass Index)**: A calculated value based on height and weight, used to classify underweight, normal weight, overweight, or obesity. [29]
- **CLASS**: The classification label for the patient, which could indicate.
  - **N (Normal)** – The patient has no diagnosed metabolic disorders (i.e., no diabetes, obesity, or other related conditions).

  - **P (Prediabetic)** – The patient is at risk of developing diabetes but has not yet been diagnosed as diabetic. This could be based on slightly elevated **HbA1c** or other risk factors.

  - **Y (Diabetic/Yes)** – The patient has been diagnosed with diabetes, possibly Type 1 or Type 2, based on medical criteria like **HbA1c** levels.



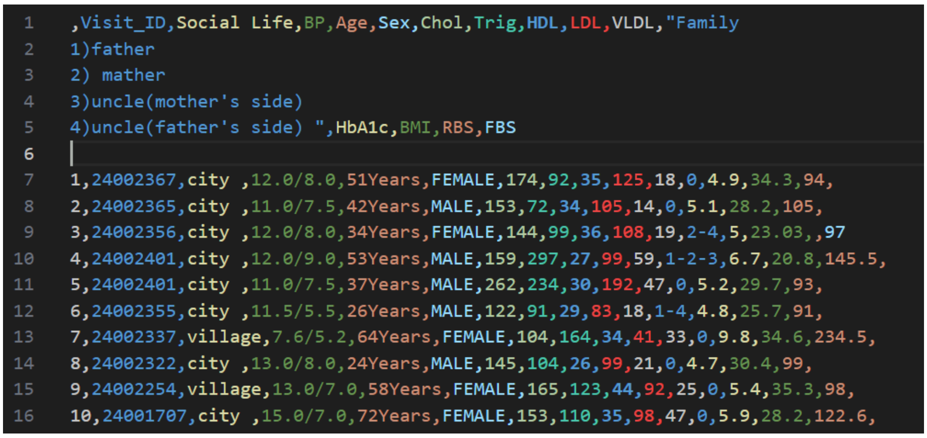
```
ID,No_Pation,Gender,AGE,Urea,Cr,HbA1c,Chol,TG,HDL,LDL,VLDL,BMI,CLASS
146,45368,M,30,6,97,5.8,4.2,1.7,1.2,2.2,0.8,19,P
152,45378,M,39,5,106,6.4,3.7,2,0.8,2.1,0.9,19.5,P
178,45384,F,48,5.6,79,6.3,4.2,1.2,2.5,2.7,1.4,27,P
12,23975,M,31,3,60,12.3,4.1,2.2,0.7,2.4,15.4,37.2,Y
18,23977,M,30,7.1,81,6.7,4.1,1.1,1.2,2.4,8.1,27.4,Y
24,23979,M,45,4.1,63,10.2,4.8,1.3,0.9,3.3,9.5,34.3,Y
675,33656789,M,45,4.1,63,10.2,4.8,1.3,0.9,3.3,9.5,34.3,Y
39,23984,M,45,5.3,77,11.2,3.9,1.5,1.3,2,10.4,29.5,Y
```

Figure 4: Sample of Diabetes Dataset

Erbil Diabetes Dataset [34]

Collected in Erbil, Kurdistan Region of Iraq in 2024, this dataset contains data from individuals referred by expert physicians for diabetes-related tests due to suspicions of the disease. The tests were conducted by highly trained professionals in a private laboratory in Erbil. The dataset comprises 661 observations with 14 features:

- Social Life: does the patient reside in an urban area (city) or a rural area (village).
- Blood Pressure (BP): The patient's blood pressure measurement.
- Age
- Sex: gender of the patient
- Cholesterol Level (Chol): The total cholesterol level in the patient's blood. [32]
- Triglycerides (Trig): The level of triglycerides in the patient's blood. [28]
- High-Density Lipoprotein (HDL): The level of HDL "good" cholesterol. [26]
- Low-Density Lipoprotein (LDL): The level of LDL "bad" cholesterol. [27]
- Very Low-Density Lipoprotein (VLDL): The level of VLDL cholesterol, another type of 'bad' cholesterol. [33]
- Family Relationship (Genetics): Reflects whether the patient has a family history of diabetes, indicating genetic susceptibility. [35]
- Glycated Hemoglobin (HbA1c): A measure of average blood glucose levels over the last two to three months, often used as a diagnostic marker. [24]
- Body Mass Index (BMI): A calculation based on height and weight used to assess body fat levels.[29]
- Random Blood Sugar (RBS): The glucose level in the blood at a randomly chosen time of day. [36]
- Fasting Blood Sugar (FBS): The blood sugar level after fasting, typically measured after not eating for at least 8 hours. [36]



```
,Visit_ID,Social Life,BP,Age,Sex,Chol,Trig,HDL,LDL,VLDL,"Family
1)father
2) mather
3)uncle(mother's side)
4)uncle(father's side) ",HbA1c,BMI,RBS,FBS

1,24002367,city ,12.0/8.0,51Years,FEMALE,174,92,35,125,18,0,4.9,34.3,94,
2,24002365,city ,11.0/7.5,42Years,MALE,153,72,34,105,14,0,5.1,28.2,105,
3,24002356,city ,12.0/8.0,34Years,FEMALE,144,99,36,108,19,2-4,5,23.03,,97
4,24002401,city ,12.0/9.0,53Years,MALE,159,297,27,99,59,1-2-3,6.7,20.8,145.5,
5,24002401,city ,11.0/7.5,37Years,MALE,262,234,30,192,47,0,5.2,29.7,93,
6,24002355,city ,11.5/5.5,26Years,MALE,122,91,29,83,18,1-4,4.8,25.7,91,
7,24002337,village,7.6/5.2,64Years,FEMALE,104,164,34,41,33,0,9.8,34.6,234.5,
8,24002322,city ,13.0/8.0,24Years,MALE,145,104,26,99,21,0,4.7,30.4,99,
9,24002254,village,13.0/7.0,58Years,FEMALE,165,123,44,92,25,0,5.4,35.3,98,
10,24001707,city ,15.0/7.0,72Years,FEMALE,153,110,35,98,47,0,5.9,28.2,122.6,
```

Figure 5: Sample of Erbil Diabetes Dataset

References

WEEKS 1 & 2

[1] Yahe Yang, Chengyue Huang, 6th of June, 2025, “A Tree-based RAG-Agent Recommendation System: A Case Study in Medical Test Data”, arXiv:2501.02727v1

[2]Ofir Ben Shoham, Nadav Rappoport, 2nd of May, 2024, ‘’ CPLLM: CLINICAL PREDICTION WITH LARGE LANGUAGE MODELS’’, https://arxiv.org/pdf/2309.11295

[3] Yogesh Kumar, Apeksha Koul, Ruchi Singla, Muhammad Fazal Ijaz, 13th of January, 2022, ”Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda”, https://link.springer.com/article/10.1007/S12652-021-03612-Z

WEEK 3

[4] Peruvian Ministry of Health, (May 2024) *Government-Backed Blood Test Analysis for Disease Prediction in Peru*. https://datosabiertos.gob.pe

[5] Meng, N., Zhang, P., Li, J., He, J., & Zhu, J. (2018). *Prediction of Coronary Heart Disease Using Routine Blood Tests*. https://arxiv.org/abs/1809.09553

[6] Jin M., Yu Q., Shu D., Zhang C., Zhu S. (Mar 2024). *Health-LLM: Personalized Retrieval-Augmented Disease Prediction System.* https://arxiv.org/html/2402.00746v5

[7] Zelin C., Chung Wendy K., Jeanne M., Zhang G., Weng C. (Sep 2024). *Rare disease diagnosis using knowledge guided retrieval augmentation for ChatGPT.* https://www.sciencedirect.com/science/article/abs/pii/S1532046424001205

[8] M. A. K. Raiaan et al., "A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges," in IEEE Access, vol. 12, pp. 26839-26874, 2024, doi: 10.1109/ACCESS.2024.3365742. https://ieeexplore.ieee.org/document/10433480

[9] Tianqi Chen, Carlos Guestrin, (June 2016). “XGBoost: A Scalable Tree Boosting System”, arXiv:1603.02754v3

[10] (Jan 2025) “What are AI Hallucinations?” https://www.k2view.com/what-are-ai-hallucinations/

[11] https://globalgenes.org/rare-disease-facts/

[37] Vanna Winland, Erika Russi (Aug 2024) "What is LlamaIndex?" https://www.ibm.com/think/topics/llamaindex

[38] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, Hervé Jégou, (Feb 2025) "The Faiss library" arXiv:2401.08281v3

WEEK 4

[12] Martínez-deMiguel C, Segura-Bedmar I, Chacón-Solano E, Guerrero-Aspizua S. (Jan 2022) "The RareDis corpus: A corpus annotated with rare diseases, their signs and symptoms." https://pubmed.ncbi.nlm.nih.gov/34879250/

[13] Santos-Silva M. A., Sousa N. , Sousa J. C., Artificial intelligence in routine blood tests, Frontiers in Medical Engineering, VOLUME 2, 2024

[14] Shams UA, Javed I, Fizan M, Shah AR, Mustafa G, Zubair M, Massoud Y, Mehmood MQ, Naveed MA. Bio-net dataset: AI-based diagnostic solutions using peripheral blood smear images. Blood Cells Mol Dis. 2024 Mar;105:102823. doi: 10.1016/j.bcmd.2024.102823. Epub 2024 Jan 4. PMID: 38241949.

[15] Personal Health Data Protection Law, 2021, (https://www.kvkk.gov.tr/Icerik/7137/2021-761), 2021/761.

[16] The sharing of health data with third parties, 2018, (https://www.kvkk.gov.tr/Icerik/5364/2018-143), 2018/143.

[17] Obligation to protect personal health data in Turkish law, 2022, (https://dergipark.org.tr/en/download/article-file/2523823)

[18] Van Panhuis, W.G., Paul, P., Emerson, C. et al. A systematic review of barriers to data sharing in public health. BMC Public Health 14, 1144 (2014). https://doi.org/10.1186/1471-2458-14-1144

[19] Dove, E.S., Phillips, M. (2015). Privacy Law, Data Sharing Policies, and Medical Data: A Comparative Perspective. In: Gkoulalas-Divanis, A., Loukides, G. (eds) Medical Data Privacy Handbook. Springer, Cham. https://doi.org/10.1007/978-3-319-23633-9_24

[20] El Emam, K., Mosquera, L., Fang, X., & El-Hussuna, A. (2022). Utility metrics for evaluating synthetic health data generation methods: Validation study. JMIR medical informatics, 10(4), e35734.

[21] Murtaza, H., Ahmed, M., Khan, N. F., Murtaza, G., Zafar, S., & Bano, A. (2023). Synthetic data generation: State of the art in health care domain. Computer Science Review, 48, 100546.

[22] Hernandez-Matamoros, A., Fujita, H., & Perez-Meana, H. (2020). A novel approach to create synthetic biomedical signals using BiRNN. Information Sciences, 541, 218-241.

WEEK 5

[23] Ahlam Rashid, 16 July 2020, Diabetes Dataset - Mendeley Data

[24] Emily Eyth; Roopa Naik., March 13 2023, Hemoglobin A1C - StatPearls - NCBI Bookshelf
[25] Dataset obtained through collaboration with Üsküdar Diabetes and Obesity Center, Fatih Sultan Mehmet Education and Research Hospital

[26] Adrian Bailey; Shamim S. Mohiuddin., 26 September 2022, Biochemistry, High Density Lipoprotein - StatPearls - NCBI Bookshelf

[27] Yasaman Pirahanchi; Hadeer Sinawe; Manjari Dimri., 8, August 2023, Biochemistry, LDL Cholesterol - StatPearls - NCBI Bookshelf

[28] High Blood Triglycerides | NHLBI, NIH

[29] https://www.nhlbi.nih.gov/health/educational/lose_wt/BMI/bmi-m.htm

[30] Thu Nguyen, Rabindra Khadka, Nhan Phan, Anis Yazidi, Pål Halvorsen, Michael A. Riegler, 16 May 2023, Combining datasets to increase the number of samples and improve model fitting https://arxiv.org/pdf/2210.05165

[31] Adrian O. Hosten. BUN and Creatinine - Clinical Methods - NCBI Bookshelf

[32] Blood Cholesterol - Diagnosis | NHLBI, NIH

[33] Fernando M. Juarez Casso; Khashayar Farzam. 22 December 2022, Biochemistry, Very Low Density Lipoprotein - StatPearls - NCBI Bookshelf

[34] Salar Amin Raheem, Amal Taha, Ibrahim Ismael Hamarash, 14 August 2024, Erbil Diabetes Dataset - Mendeley Data

[35] Daniel David, Joanne Dalton, Cherlie Magny-Normilus, Maura Moran Brain, Tyler Linster, Sei J Lee, May 2019, The Quality of Family Relationships, Diabetes Self-Care, and Health Outcomes in Older Adults, https://pmc.ncbi.nlm.nih.gov/articles/PMC6528399/

[36] Blood Glucose Test - Test Overview

***OBSERVATION**: References numbers 37 and 38 are in the WEEK 3 section.*