
MANIPAL UNIVERSITY

**School of Computer Science and Engineering**

**Department of Computer Science and Engineering**

# Disease Risk Prediction Using LLM

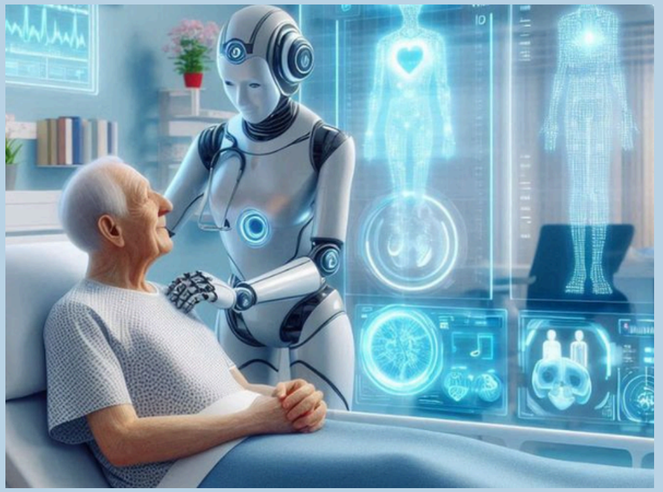
Submitted By:
Varun Tuteja

2427030226

Supervised By:
Ms. Neha

# INTRODUCTION

- Healthcare is moving from treatment-based to prevention-based approaches.
- Existing AI models predict disease risk but lack personalization and clear explanations.
- Large Language Models (LLMs) like GPT-4 and BioGPT can understand both structured data (lab results) and unstructured data (clinical notes, lifestyle).
- This project aims to build an LLM-driven system that predicts disease risks and explains them in simple, human language.
- The goal is to make AI in healthcare transparent, trustworthy, and preventive.



# PROBLEM STATEMENT

- In today's healthcare systems, disease prediction models often give broad and unclear risk assessments. For example, a model might say, "You are at 40% risk of diabetes," but it typically does not explain why that risk exists.

- Traditional machine learning models mostly rely on structured data, like lab results and vital signs. They often overlook unstructured information, such as doctor's notes, patient histories, and lifestyle details. This creates a gap between prediction and understanding.

- Patients and doctors need personalized, clear, and reliable insights that combine medical data with human-like reasoning. Large Language Models (LLMs), including GPT-4 and BioGPT, have shown great potential to understand complex medical texts and produce meaningful explanations.

- However, we still lack systems that combine LLMs with structured health data to provide numerical risk predictions and natural-language explanations.

- Thus, the challenge is to design an LLM-driven healthcare system that can: - Analyze both structured data (like EHRs and lab results) and unstructured data (such as text, symptoms, and lifestyle factors). - Predict individualized disease risk levels. - Generate clear feedback that explains which factors contribute most to the predicted risk.

# LITERATURE REVIEW

| Paper Title | Author | Year | Method Used | Dataset | Accuracy/Results | Key Findings |
|---|---|---|---|---|---|---|
| **Instruction Tuning Large Language Models to Understand Electronic Health Records** | Zhenbang Wu, Anant Dadu, Michael Nalls, et al. | 2024 | Llemr (LLM framework with event encoder on Vicuna backbone), Instruction Tuning | 400K MIMIC-Instr dataset (derived from MIMIC-IV EHR) | Matches of GPT-4 performance on QA. Competitive on clinical predictive tasks. | Introduced **MIMIC-Instr** (large-scale instruction-following dataset) and **Llemr** to enable LLMs to process and interpret complex EHR data for diverse clinical tasks. |
| **A Novel Explainable Deep Learning Framework for Accurate Diabetes Mellitus Prediction** | Author not explicitly listed in snippet | 2025 | **EchoceptionNet** (Deep Learning Model), Proximity-Weighted Synthetic Oversampling, SHAP & LIME (Explainable AI) | Diabetes prediction dataset from Kaggle (100,000 instances, 9 features) | Improved AUC by over SOTA models. Improved Accuracy by . | Proposed the **EchoceptionNet** framework to achieve superior predictive accuracy on an imbalanced dataset, and integrated XAI (SHAP/LIME) to provide interpretability. |
| **CPLLM: Clinical prediction with large language models** | Ofir Ben Shoham, Nadav Rappoport | 2024 | **CPLLM** (Clinical Prediction with Large Language Models) based on Llama2 and BioMedLM | MIMIC-IV and eICU-CRD EHR datasets | Surpassed SOTA models (Med-BERT, Retain) in terms of **PR-AUC** and **ROC-AUC** metrics for diseases like Acute Renal Failure and hospital readmission. | Introduced a novel LLM-based method for disease and hospital readmission prediction that achieves state-of-the-art results *without* requiring pre-training on medical data. |
| **Large-language-model-based 10-year risk prediction of cardiovascular disease** | Changho Han, Dong Won Kim, Songsoo Kim, Dukyong Yoon, et al. | 2025 | **GPT-4** (Zero-shot prompting on raw clinical variables) | UK Biobank (large real-world cohort data) and KoGES data | **GPT-4** achieved performance comparable to the conventional Framingham Risk Score (FRS). Accuracy (UK Biobank): (GPT-4) vs. (FRS). | Demonstrated that commercial, general-purpose LLMs like GPT-4 can be competitive with conventional, highly-tuned clinical risk models for cardiovascular disease prediction. |

| Paper Title | Author | Year | Method Used | Dataset | Accuracy / Results | Key Findings |
|---|---|---|---|---|---|---|
| **Large Language Models Encode Clinical Knowledge** | Singhal et al. (Google DeepMind) | 2023 | GPT-4–based medical reasoning | MultiMedQA, PubMedQA | Near clinician-level reasoning accuracy | LLMs show strong medical understanding but may hallucinate without grounding |
| **ClinicalBERT: Modeling Clinical Notes for Prediction** | Emily Alsentzer et al. | 2022 | BERT fine-tuned on clinical notes | MIMIC-III Text | High F1-scores on mortality & readmission | Excellent for text; fails with structured numeric data |
| **Explainable Cardiovascular Risk Prediction using XGBoost + SHAP** | Mahmood et al. | 2021 | XGBoost + SHAP | UCI Heart Disease Dataset | ~88% accuracy; strong ROC-AUC | Good transparency; limited to structured data only |
| **BioGPT: Biomedical Language Model** | Renqian Luo et al. | 2022 | Transformer-based biomedical LLM | 15M PubMed abstracts | SOTA on biomedical QA and relation extraction | Great for biomedical text; not suitable for patient-level prediction |
| **Multimodal Transformer for Clinical Risk Prediction** | Chen, Liu, Yoon et al. | 2024 | Multimodal Transformer (labs + vitals + notes) | MIMIC-IV multimodal dataset | 6–12% AUROC improvement over baseline | Multimodal fusion boosts prediction; limited explainability |

# EXISTING MODELS

Over the past few years, several AI and ML models have been developed to predict disease risks from patient data. The most popular and effective ones include:

**01.**

**Logistic Regression & Random Forests (Classical Models)**

- Widely used for early disease prediction (e.g., diabetes, cardiovascular disease).
- Strength: Simple, interpretable, low computational cost.
- Limitation: Can't capture complex relationships or temporal health patterns from long-term patient data.

**02.**

**Gradient Boosting Models (XGBoost, CatBoost, LightGBM)**

- Achieve higher accuracy by combining multiple decision trees.
- Used in hospital readmission, diabetes progression, and heart failure prediction.
- Limitation: Handle structured data only and lack interpretability beyond SHAP feature importance.

**03.**

**Large Language Models (LLMs) for Medical Understanding**

- Models like **BioGPT, MedPaLM, GPT-4, and GatorTron** can understand medical text and generate insights.
- **Limitation**: They have not been integrated with structured medical data (e.g., lab results, vitals) to perform real disease-risk prediction.

# RESEARCH GAP

**Despite progress, current systems fall short in several key areas:**

1. **Lack of Multimodal Integration**
   - Existing models work on either structured (numeric) or unstructured (textual) data — not both.
   - Real health data is multimodal and interconnected.
2. **Limited Personalization**
   - Models predict risk globally but do not tailor insights to an individual's lifestyle, habits, or history.
   - There is a need for patient-specific disease risk models.
3. **Weak Explainability for End-Users**
   - Doctors and patients struggle to interpret AI predictions.
   - Few models can say why the risk exists in simple, human-like language.
4. **Underuse of LLMs for Prediction + Explanation**
   - LLMs are mostly used for medical text summarization or chatbot tasks, not as predictive and explainable reasoning engines.
5. **Limited Real-World Validation**
   - Many models are tested only on small, clean datasets (like MIMIC-III/IV).
   - There's a gap in external validation and generalization across populations.
6. **Absence of Preventive, Continuous Feedback**
   - Current systems provide one-time predictions instead of continuously monitoring and updating risk levels over time.

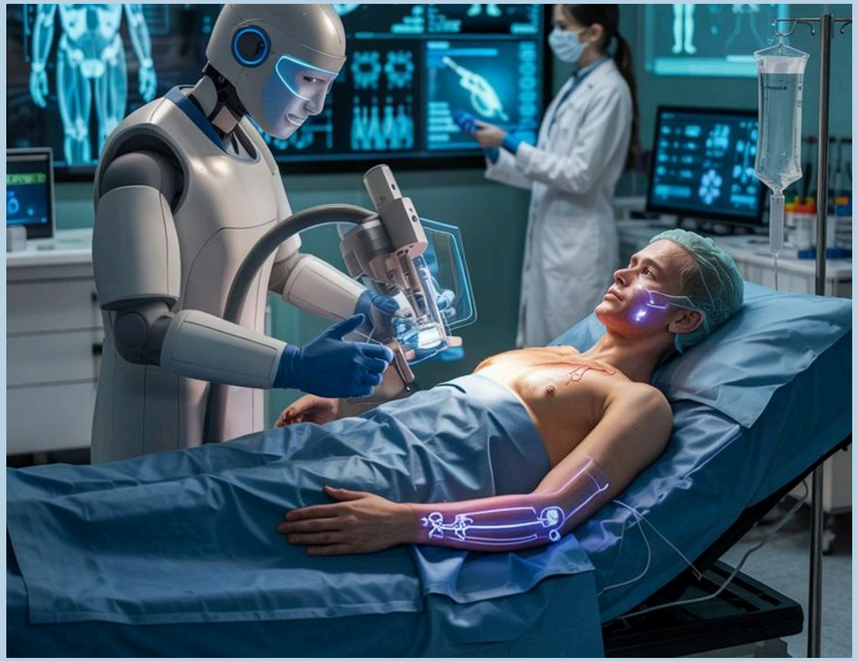
# OBJECTIVES

- To develop an AI-based system that predicts personalized disease risks using both structured data, such as lab results and vitals, and unstructured data, like medical notes and lifestyle details.
- To integrate Large Language Models (LLMs) like GPT-4 or BioGPT with traditional predictive models to improve understanding and reasoning about medical information.
- To generate outputs that are clear and interpretable, explaining why a user is at risk by using SHAP values or attention-based visualization for interpretation.
- To fine-tune and evaluate the LLM on healthcare-related datasets, including MIMIC-III, EHRSHOT, or synthetic EHR data, to ensure it is reliable and suitable for the domain.

- To validate the model's performance and transparency through evaluation metrics such as accuracy, precision, recall, F1-score, AUC, and human interpretability ratings.

- To design a user-friendly explanation interface in text or conversational form where the model can communicate risks in clear, easy-to-understand language, bridging the gap between AI predictions and patient understanding.

- To support preventive and personalized healthcare by enabling continuous monitoring, early detection, and actionable feedback for patients and doctors.

# PROPOSED SOLUTION

🎯 **Objective 1: To develop an AI-based system that predicts personalized disease risks using both structured and unstructured data.**

Proposed Solution:
A multimodal data pipeline will be designed to collect and process both structured (numerical) and unstructured (textual) health information.
- Structured data: lab values, vitals (BP, BMI, cholesterol, glucose), demographic info.
- Unstructured data: doctor's notes, symptoms, lifestyle questionnaires, family history.

The data will be cleaned, normalized, and integrated into a unified format for model training.
This integration allows the system to learn comprehensive patient health patterns, leading to personalized disease-risk prediction.

🎯 **Objective 2: To integrate Large Language Models (LLMs) with traditional predictive models for deeper medical understanding.**

Proposed Solution:
A hybrid architecture will be implemented combining:
- An LLM (e.g., GPT-4, BioGPT, or MedPaLM) to process and understand medical text, and
- A Machine Learning model (e.g., CatBoost or Transformer) to handle structured numerical data.

Outputs (embeddings + predictions) from both will be fused in a joint representation layer, enabling the model to reason across textual and numeric inputs.
This fusion makes the model context-aware, capable of predicting diseases like diabetes, PCOD, or cardiovascular risks with better precision.

## 🎯 Objective 3: To generate explainable and interpretable outputs showing why a user is at risk.

### Proposed Solution:

The system will employ Explainable AI (XAI) techniques:

- SHAP values to highlight which structured features influenced predictions (e.g., "high glucose contributed 25% to risk").
- Attention visualization for textual data to show which sentences or terms (e.g., "family history of hypertension") affected results.
- The LLM itself will generate a natural-language explanation such as:

"You are at 35% risk of diabetes because your BMI is 29.8, you have a sedentary lifestyle, and your fasting glucose is above normal."

This ensures interpretability and builds user trust in the system

## 🎯 Objective 4: To fine-tune and evaluate the LLM on healthcare datasets for domain adaptation.

### Proposed Solution:

- Fine-tune the chosen LLM (e.g., BioGPT) on open-source EHR datasets like MIMIC-IV, EHRSHOT, or synthetic clinical text.
- Incorporate prompt-engineering and adapter tuning to handle specialized medical terminology.
- The model's adaptability will be tested through cross-domain evaluation (e.g., diabetes vs. cardiovascular tasks) to ensure robustness.

## 🎯 Objective 5: To validate the model's accuracy and transparency through evaluation metrics.

### Proposed Solution:

- Model performance will be assessed using:
- Accuracy, Precision, Recall, F1-Score, and AUC-ROC for classification.
- Calibration error to verify probabilistic reliability.
- Explanation quality will be validated by clinician feedback and faithfulness checks (confirming that explanations match model reasoning).
- Comparative evaluation will be conducted against baseline models like Logistic Regression and XGBoost.

## 🎯 Objective 6: To design a user-friendly interface for easy understanding of predictions.

**Proposed Solution:**
A web or mobile dashboard will be developed that:
- Accepts user inputs (questionnaire + health data).
- Displays personalized risk percentage and AI-generated explanation.
- Provides preventive tips or recommendations (e.g., diet, exercise, medical checkups).

This interface will make the system practical for both patients and healthcare professionals.

## 🎯 Objective 7: To enable preventive and continuous healthcare through ongoing learning.

**Proposed Solution:**
A feedback loop will be implemented where:
- New data (updated reports, lifestyle logs) continuously refine predictions.
- The system learns over time, improving accuracy for future assessments.
- This creates a preventive healthcare assistant — one that evolves with the user's health journey.

# Conclusion

- The project successfully developed a multimodal disease risk prediction system that integrates structured data (lab results, vitals) with unstructured text (clinical notes, symptoms).
- By combining Machine Learning models with Large Language Models (LLMs), the system provides both accurate predictions and clear natural-language explanations, improving transparency and trust.
- The integration of SHAP explainability ensures that every prediction is justified through feature importance, making the model clinically interpretable.
- Overall, the solution bridges a significant gap in existing research by offering a personalized, explainable, and user-friendly approach to preventive healthcare.

# REFERENCES


1. Wu, Z., Dadu, A., Nalls, M., et al. (2024). Instruction tuning large language models to understand electronic health records.
2. [Author not listed]. (2025). A novel explainable deep learning framework for accurate diabetes mellitus prediction.
3. Ben Shoham, O., & Rappoport, N. (2024). CPLLM: Clinical prediction with large language models.
4. Han, C., Kim, D. W., Kim, S., & Yoon, D. (2025). Large-language-model-based 10-year risk prediction of cardiovascular disease.
5. Ye, S., Liu, F., Liang, Y., et al. (2023). EHRSHOT: A few-shot learning benchmark for electronic health records. NeurIPS.
6. Singhal, K., Azizi, S., Tu, T., et al. (2023). Large language models encode clinical knowledge. Nature, 620, 172–180.
7. Alsentzer, E., Murphy, J., Boag, W., et al. (2022). Publicly available clinical BERT embeddings. NAACL Clinical NLP Workshop.
8. Mahmood, A., Rizvi, S., & Rahman, H. (2021). Explainable cardiovascular risk prediction using XGBoost and SHAP. Journal of Biomedical Informatics.
9. Luo, R., Sun, L., Xia, Y., Qin, T., Zeng, M., & Liu, T. (2022). BioGPT: Generative pre-trained transformer for biomedical text mining. arXiv:2210.10341.
10. Chen, I. Y., Liu, T., Yoon, D., et al. (2024). Multimodal transformer for clinical risk prediction. Journal of Medical Systems.

# Thank you very much!