Week 9: Select appropriate AI Models

This document presents the research and selection of appropriate Artificial Intelligence (AI) models for the **Hypotify Clinical Insights Bot**. The selected models—a **Text-to-SQL Language Model** and a **Gradient Boosting Classifier**—are designed to leverage the highly structured relational EMR database developed in ITEC 5020 and fulfill the dual purpose of the virtual assistant.

Select Appropriate AI Models

Abstract

The **Hypotify Clinical Insights Bot** requires integrating two distinct AI model types to achieve its dual functionality: natural language querying and predictive analysis. This paper researches and selects a **Text-to-SQL Model** (based on the Bidirectional Encoder Representations from Transformers [BERT] architecture) for translating user input into MySQL queries, and a **Gradient Boosting Machine (GBM)** for running predictive risk classifications on the structured EMR data. The selection is justified based on the models' suitability for highly structured clinical data, interpretability requirements in healthcare, and the availability of robust open-source Python libraries (Scikit-learn and Hugging Face) for implementation in the subsequent ITEC 5025 course. The paper concludes that this dual-model architecture provides the optimal balance of user convenience and computational reliability necessary for clinical informatics.

Introduction

The successful development of a database for an AI virtual assistant requires that the underlying data architecture (the 3NF MySQL schema) is perfectly aligned with the computational needs of the integrated AI models. The **Hypotify Clinical Insights Bot** operates within the Healthcare domain, utilizing a structured, relational database populated with synthetic EMR data. This structure dictates that the chosen

AI models must be highly efficient at consuming numerical and categorical data for prediction, and robust at translating complex human language into precise SQL queries.

This paper outlines the research and selection process for the final AI architecture, which is broken down into two essential, domain-specific tasks. The final selections prioritize models known for high performance on tabular data and models capable of producing transparent, grounded outputs, which is a key ethical requirement in clinical applications (Rajkomar, Dean, & Kohane, 2019).

Task 1: Research and Selection of the Interface Model

The first core task of the Hypotify Clinical Insights Bot is providing a seamless user interface that allows researchers to query the database using natural language rather than writing complex SQL code.

Research Appropriate AI Models

This task falls under **Natural Language Processing (NLP)**, specifically the **Text-to-SQL** problem.

- **Rule-Based Systems:** These rely on defining every possible keyword and grammatical rule to build SQL queries. They offer high control but are rigid, fragile, and fail easily when faced with slightly varied inputs (e.g., confusing "high glucose" with "glucose level high").

- **Sequence-to-Sequence (Seq2Seq) Models (Transformers):** Modern solutions utilize transformer models (like BERT or T5) that are pre-trained on massive amounts of text and then fine-tuned on specialized datasets (like WikiSQL or Spider) to learn the grammatical patterns required to map a natural language question (the sequence) to an SQL statement (the new sequence). These models are highly adaptable to variations in user language.

Selected Model: Fine-Tuned BERT/LLM (Text-to-SQL Model)

The selected approach is a **Fine-Tuned Transformer-based Language Model** (e.g., leveraging the architecture of **BERT**).

Justification for Suitability

1. **Robustness and Accuracy:** Text-to-SQL models based on transformers are state-of-the-art for semantic parsing. They can handle the variety of phrasing researchers might use to ask for the same clinical cohort (e.g., "patients with high blood sugar" vs. "glucose greater than 200"). This level of robustness is impossible with fragile, rule-based systems.

2. **Integrity Grounding:** The model's output is not conversational; it is a **validated SQL string**. This forces the chatbot to rely entirely on the structured data in the MySQL database, ensuring the generated content (the patient cohort) is accurate and grounded in the established 3NF schema, mitigating the risk of factual errors often associated with purely generative chat interfaces.

3. **Schema Awareness:** Transformer-based models can be fine-tuned specifically on the schema of the target database (PATIENT, ADMISSION, LAB_OBSERVATION), making them "schema-aware" and ensuring they generate valid column names and join paths necessary for the highly normalized relational model.

Task 2: Research and Selection of the Predictive Model

The second core task is generating actionable clinical insights by running risk predictions on the retrieved patient cohorts.

Research Appropriate AI Models

This task falls under **Supervised Machine Learning**, specifically **Classification** (predicting a binary or categorical outcome, such as "High Risk" or "Low Risk" for a disease or complication). The EMR data is highly tabular, featuring numerical (lab values, age) and categorical (race, gender) features.

- **Logistic Regression:** Simple, highly interpretable, but often lacks the predictive power needed for complex, multi-variable clinical outcomes.

- **Support Vector Machines (SVMs):** Good for complex boundaries but can be slow to train on large datasets and less interpretable.

- **Random Forest (RF):** An ensemble method that handles non-linear relationships well, runs quickly, and provides reasonable feature importance (interpretability).

- **Gradient Boosting Machines (GBMs) / XGBoost:** An advanced ensemble method known for achieving state-of-the-art performance on structured, tabular datasets like those found in EMRs.

Selected Model: Gradient Boosting Machine (GBM)

The selected model is a **Gradient Boosting Machine (GBM)**, such as **XGBoost** or **LightGBM**.

Justification for Suitability

1. **High Performance on Tabular Data:** GBMs consistently outperform deep learning models and simpler ensemble methods (like Random Forest) when dealing with complex, non-image, and non-text structured data (Rajkomar et al., 2019). Given the EMR data consists of discrete columns (lab values, admission dates, demographics), a GBM is the optimal choice for predictive power.

2. **Feature Importance and Interpretability:** In healthcare, models must not only be accurate but also **explainable**. GBMs provide clear feature importance scores (e.g., showing that "Glucose Value" is 10x more important than "Patient Language" in predicting diabetes risk). This aligns with ethical guidelines for transparency in clinical AI decision-making.

3. **Handles Complexity:** The EMR database contains mixed data types (numerical lab values, categorical race/gender, time-series elements). GBMs naturally handle this combination without extensive feature engineering or scaling required by models like Neural Networks (Kohn & Alisic, 2018).

Exploration of Pre-trained Models and Libraries

The ITEC 5025 development process will utilize readily available and robust open-source Python libraries, avoiding the need to train complex models from scratch where possible.

Natural Language Processing (Text-to-SQL)

- **Libraries: Hugging Face Transformers** and **simple LLM APIs** (if applicable).
- **Availability:** While pre-trained general Text-to-SQL models exist (e.g., those fine-tuned on the Spider dataset), they are not schema-aware of the artificial_emr database. Therefore, the implementation will involve either:
    - **Prompt Engineering/Function Calling:** Using a simpler LLM API (like a specialized Gemini model) and providing the EMR schema definitions as context, letting the model generate the SQL via function calling—a modern, efficient approach.
    - **Transfer Learning:** Fine-tuning a smaller, pre-trained BERT model (available via Hugging Face) using a synthetic dataset generated specifically from the artificial_emr schema to teach it the database structure.

Machine Learning (Predictive Classification)

- **Libraries: Scikit-learn** and **XGBoost/LightGBM**.

- **Availability:** These are open-source, highly optimized Python libraries. Training the final classification model will be done entirely on the structured data retrieved directly from the **artificial_emr** database using a straightforward Scikit-learn or XGBoost pipeline. No external pre-trained model is required, as the input features (lab values, demographics) are unique to this EMR structure.

Conclusion

The selection of a **Fine-Tuned Text-to-SQL Model** and a **Gradient Boosting Machine (GBM)** creates a robust, two-tier AI architecture for the Hypotify Clinical Insights Bot. The selection of these models directly reflects the structured nature of the 3NF MySQL database, ensuring that the AI is not only fast and accurate (thanks to the GBM's performance on tabular data) but also accountable and reliable (thanks to the grounded output of the Text-to-SQL model). The availability of industry-standard libraries like Hugging Face and Scikit-learn ensures that the implementation in ITEC 5025 will be practical and achievable.

References

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G. S., Thrun, S., & Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine*, *25*(1), 24–29.

Kohn, T. A., & Alisic, M. A. (2018). Machine learning methods for analyzing clinical and administrative data: A systematic review. *Journal of Biomedical Informatics*, *87*, 136-146.

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in health care. *The New England Journal of Medicine*, *380*(14), 1347–1358.

Zhang, Y., Qiu, M., Tsai, C., Hassan, M. M., & Alamri, A. (2017). Health-CPS: Healthcare Cyber-Physical System Assisted by Cloud and Big Data. *IEEE Systems Journal*, *11*(1), 88–95.