VITAI: An Enhanced Deep Learning Health Severity Index Using Advanced Data Preprocessing and Integration of Structured and Unstructured EHR Data

---

Introduction

The integration of Electronic Health Records (EHRs) into healthcare systems has unlocked vast potential for data-driven decision-making and advanced analytics. However, privacy concerns and the complexity of processing diverse data types hinder the full realisation of this potential. **VITAI** proposes a solution by developing a health severity index ranging from 0 (healthy) to 10 (most severe), powered by deep learning techniques that effectively process both structured data and freely typed clinical notes. By employing advanced data preprocessing methods and integrating autoencoders with transformer models, VITAI aims to deliver predictive analytics that aid real-time decision-making while preserving patient privacy through the use of synthetic healthcare data.

Aims, Objectives, and Scope

Aim

VITAI aims to develop an accurate and interpretable health severity index by integrating structured and unstructured EHR data using advanced data preprocessing techniques and deep learning models such as autoencoders and transformers. This tool will provide healthcare professionals with actionable insights, facilitating informed decision-making based on the patient's predicted health severity level.

Objectives

- **Advanced Data Preprocessing**
    - Implement robust data preprocessing techniques for structured data, including handling missing values, one-hot encoding of categorical variables, and appropriate feature scaling to optimise model performance.
    - Ensure alignment between data preprocessing steps and model requirements, such as matching activation functions with data scaling methods in the autoencoder.
- **Autoencoder Development**
    - Design and train autoencoders with adjusted architectures to capture essential patterns in structured EHR data, enhancing the model's ability to learn meaningful representations.

    - Address data preprocessing issues that affect model performance, such as scaling input features to the [0, 1] range when using sigmoid activation functions.
- **Transformer Integration**
    - Utilise transformers to process and understand unstructured clinical notes, extracting critical health risks and contextual information hidden in free-text entries.
    - Fine-tune transformer models to handle domain-specific language in clinical texts effectively.
- **Health Severity Index Generation**
    - Combine insights from both models to generate an accurate health severity index, providing a transparent and interpretable score from 0 to 10.
    - Implement a method to integrate outputs from the autoencoder and transformer, possibly through clustering techniques or additional neural network layers.
- **Explainable AI Interface**
    - Develop an explainable AI (XAI) interface to ensure transparency in predictions and support decision-making processes in healthcare environments.
    - Use XAI tools to highlight how different data features contribute to the health severity score.

Scope

The project focuses on building an integrated deep learning system that processes structured and unstructured EHR data, emphasising advanced data preprocessing techniques to improve model performance. The autoencoders will handle structured data with adjusted architectures and preprocessing steps, while transformers will process free-text clinical notes. The scope includes data generation, preprocessing, model development, testing, and a user interface for real-time interaction with healthcare providers.

Definition of the Problem

Healthcare professionals often face challenges in processing vast amounts of patient data, especially when clinical notes are freely typed and unstructured. Traditional decision-making tools may overlook the unstructured nature of these notes and suffer from inadequate data preprocessing, leading to reduced predictive power. Moreover, improper handling of categorical variables and scaling issues in structured data can hinder model performance. **VITAI** addresses these gaps by integrating advanced data preprocessing techniques and a transformer-based model that

processes free-text clinical notes, extracting essential information. Combined with a well-tuned autoencoder for structured data, this approach enhances health severity predictions, helping healthcare professionals make more informed decisions and identify patient risks early.

Background Review

- **Existing Models**
    - **Autoencoders**: Widely used for structured data analysis, but their effectiveness is highly dependent on proper data preprocessing and model architecture adjustments.
    - **Transformers**: Models like BERT and GPT have revolutionised natural language processing, excelling at extracting insights from unstructured text.
- **Gaps in Current Solutions**
    - **Data Preprocessing Challenges**: Inadequate handling of categorical variables and scaling in structured data can lead to suboptimal model performance.
    - **Integration Difficulties**: Existing health severity prediction models often struggle to effectively combine insights from structured and unstructured data sources.

VITAI aims to bridge these gaps by implementing advanced data preprocessing techniques and integrating both autoencoders and transformers to improve decision-making accuracy significantly.

Project Architecture

VITAI consists of several interlinked modules covering data generation, advanced data preprocessing, model training, and prediction generation, integrating structured and unstructured data.

1. Data Generation Module

- **Sources**: Synthetic EHR data from repositories like **Synthea** and **NHSX Skunkworks**.
- **Function**: Generate realistic patient data, including structured information (e.g., vitals, labs) and freely typed clinical notes.

2. Advanced Data Preprocessing Module

- **Function**:

    - **Structured Data**:
        - **Data Cleaning**: Handle missing values using appropriate imputation techniques.
        - **Categorical Encoding**: Use one-hot encoding for nominal variables to avoid introducing ordinal relationships.
        - **Feature Scaling**: Apply scaling methods like **MinMaxScaler** to align with activation functions in the autoencoder.
        - **Data Type Consistency**: Ensure all features are of compatible data types (e.g., converting boolean to numeric).
    - **Unstructured Data (Clinical Notes)**:
        - **Text Preprocessing**: Tokenization, stopword removal, and lemmatization.
        - **Data Formatting**: Prepare text data for input into transformer models.
- **Tools**: **Pandas**, **NumPy**, **Scikit-learn** for structured data; **NLTK** or **SpaCy** for natural language processing.

3. Autoencoder Model Module

- **Function**:
    - Handle structured data with an adjusted architecture, incorporating changes like increased depth and width.
    - Align activation functions with data scaling (e.g., using sigmoid activation with data scaled to [0, 1]).
    - Capture key health patterns effectively through enhanced model capacity.
- **Tools**: **TensorFlow/Keras** for deep learning model development.

4. Transformer Model Module

- **Function**:
    - Process freely typed clinical notes using transformer models (e.g., **BERT** or **GPT-based models**).
    - Extract health risk indicators and contextual information from the notes.
    - Generate embeddings that capture the semantic meaning of clinical text.
- **Tools**: **Hugging Face Transformers**, **TensorFlow/Keras**.

5. Health Severity Index Module

- **Function**:
    - Integrate insights from the autoencoder and transformer models to generate a health severity score on a 0-10 scale.
    - Provide a holistic view of patient health by combining processed structured data and clinical notes.

- **Method**:
    - Use clustering algorithms or additional neural network layers to combine embeddings and reconstruction errors.
    - Implement a weighted approach to balance contributions from structured and unstructured data.

6. Explainable AI (XAI) Module

- **Function**:
    - Offer transparency by explaining the model's predictions.
    - Highlight how both structured data and clinical notes contribute to the assigned health severity score.
- **Tools**: **LIME** or **SHAP** for generating explanations and interpreting model outputs.

7. User Interface (UI) Module

- **Function**:
    - Allow healthcare professionals to input patient data and receive real-time severity index predictions.
    - Display explanations for the predictions to ensure informed decision-making.
- **Tools**: **Flask** or **Django** for web development; **Plotly** or **Dash** for data visualisation.

Use Cases and Diagrams

Use Case 1: Comprehensive Patient Assessment

A healthcare professional inputs both structured data and freely typed clinical notes. **VITAI** processes the data using advanced preprocessing techniques and adjusted autoencoder and transformer models, generating a severity index and an explanation for the score.

Use Case 2: Enhanced Risk Identification

The system analyses the patient's clinical notes for hidden risks (e.g., casually mentioned symptoms), integrating these insights with meticulously preprocessed structured data to provide a more accurate risk assessment.

Architecture Diagram

A detailed architecture diagram will depict the flow from data input through advanced preprocessing, the transformer and autoencoder models, into the health severity index module, and finally to the user interface.

Requirements Elicitation

- **Methods**:
    - **Clinician Engagement**: Conduct surveys and interviews with healthcare professionals to gather requirements and validate the usefulness of integrating unstructured clinical notes.
    - **Feedback Loops**: Implement iterative development cycles incorporating user feedback to refine the tool.
- **Tools**:
    - **Project Management Platforms**: Use tools like **Jira** or **Trello** to track requirements and tasks.
    - **Communication Channels**: Establish regular meetings and communication with stakeholders.

Time Schedule

Task	Start Date	End	Milestone
		what are you doing? Shutup.Date	

Data Generation and Preprocessing	Initial Data Prepared
Model Development (Autoencoder and Transformer)	First Model Versions
Integration and Health Severity Index Calculation	Integrated System Ready
XAI and UI Implementation	Explainable Interface
Final Testing and Validation	Model Testing Completed
Presentation Preparation	Final Presentation Ready

References / Bibliography

- **Vaswani, A. et al. (2017)**. *Attention is All You Need*.
- **Miotto, R., et al. (2016)**. *Deep Patient: Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records*.
- **NHSX Skunkworks (2020)**. *Synthetic Health Data for Machine Learning*.
- **Devlin, J., et al. (2019)**. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*.
- **Pedregosa, F., et al. (2011)**. *Scikit-learn: Machine Learning in Python*.

Conclusion

By incorporating advanced data preprocessing techniques and adjusted model architectures, **VITAI** aims to enhance the accuracy and interpretability of health severity predictions. The integration of meticulously processed structured data and insights from unstructured clinical notes positions VITAI as a powerful tool for healthcare professionals. This approach enables more informed decision-making

and early identification of patient risks while maintaining patient privacy through the use of synthetic data.

-