**1. Dữ liệu cổ phiếu (Stock Data)**

* **Nguồn dữ liệu:**
  + Bài báo sử dụng dữ liệu từ hai sàn giao dịch chứng khoán lớn: **NYSE (New York Stock Exchange) và NASDAQ**.
  + Dữ liệu được thu thập từ các nguồn công khai, bao gồm các chỉ số thị trường và giá cổ phiếu trong một khoảng thời gian dài.
* **Các đặc trưng chính (Feature Selection):**
  + **Dữ liệu chuỗi thời gian:** Giá mở cửa, giá đóng cửa, giá cao nhất, giá thấp nhất và khối lượng giao dịch.
  + **Chỉ báo kỹ thuật:** Trung bình động (SMA, EMA), RSI (Relative Strength Index), MACD (Moving Average Convergence Divergence), và Bollinger Bands.
  + **Lợi nhuận cổ phiếu:** Lợi nhuận hàng ngày, trung bình lợi nhuận trong khoảng thời gian nhất định.
* **Tiền xử lý dữ liệu (Preprocessing):**
  + Chuẩn hóa dữ liệu giá cổ phiếu để giảm ảnh hưởng của giá trị tuyệt đối giữa các cổ phiếu.
  + Chia dữ liệu thành các cửa sổ thời gian để tạo ra các mẫu huấn luyện cho mô hình LSTM.

**2. Dữ liệu quan hệ giữa cổ phiếu (Relationship Data)**

* **Xây dựng biểu đồ quan hệ cổ phiếu (Stock Relationship Graph):**
  + Các cổ phiếu không hoạt động riêng lẻ mà có mối quan hệ với nhau dựa trên **cùng ngành, cùng chuỗi cung ứng, hoặc phụ thuộc tài chính**.
  + Bài báo sử dụng các phương pháp xây dựng đồ thị quan hệ cổ phiếu để đưa vào mô hình học sâu.
* **Nguồn dữ liệu quan hệ:**
  + **Dữ liệu ngành (Industry Sector Data):** Từ các báo cáo tài chính và phân loại ngành từ các nguồn như S&P 500 hoặc Yahoo Finance.
  + **Dữ liệu từ WikiData:** Quan hệ giữa các công ty dựa trên thông tin về sở hữu, hợp tác kinh doanh.
* **Phương pháp biểu diễn quan hệ:**
  + **Đồ thị tĩnh (Static Graph):** Các mối quan hệ giữa cổ phiếu được thiết lập dựa trên dữ liệu lịch sử.
  + **Đồ thị động (Dynamic Graph):** Mối quan hệ thay đổi theo thời gian dựa trên các sự kiện kinh tế, chính trị.
* **Áp dụng vào mô hình (Integration into Model):**
  + Biểu diễn đồ thị được sử dụng để huấn luyện **Temporal Graph Convolution (TGC)** – một thành phần chính của mô hình RSR (Relational Stock Ranking).

**3. Chiến lược đánh giá mô hình và Back-Testing**

* **Phương pháp đánh giá:**
  + Dữ liệu được chia theo **temporal splits**, nghĩa là sử dụng dữ liệu trong quá khứ để huấn luyện và dữ liệu sau đó để kiểm thử.
  + Áp dụng các kỹ thuật **cross-validation theo thời gian** để đánh giá độ ổn định của mô hình.
* **Back-Testing:**
  + Mô phỏng giao dịch thực tế bằng cách chọn các cổ phiếu có xếp hạng cao nhất theo dự đoán của mô hình và so sánh lợi nhuận thực tế.
  + Các chiến lược thử nghiệm: **Top-1 selection (chọn cổ phiếu có xếp hạng cao nhất), Top-5, Top-10 (chọn danh mục gồm nhiều cổ phiếu hàng đầu).**
  + Chỉ số đánh giá chính:
    - **Mean Squared Error (MSE):** Đánh giá độ chính xác dự đoán lợi nhuận.
    - **Mean Reciprocal Rank (MRR):** Đánh giá hiệu quả xếp hạng.
    - **Investment Return Ratio (IRR):** So sánh lợi nhuận thực tế của các mô hình.

**Kết luận**

* **Dữ liệu cổ phiếu** (giá, khối lượng, chỉ báo kỹ thuật) giúp mô hình nhận diện xu hướng giá.
* **Dữ liệu quan hệ giữa cổ phiếu** (đồ thị quan hệ ngành, tài chính) giúp mô hình cải thiện hiệu suất bằng cách khai thác thông tin liên quan giữa các cổ phiếu.
* **Phương pháp đánh giá và Back-Testing** đảm bảo rằng mô hình có thể hoạt động hiệu quả trong điều kiện thực tế.