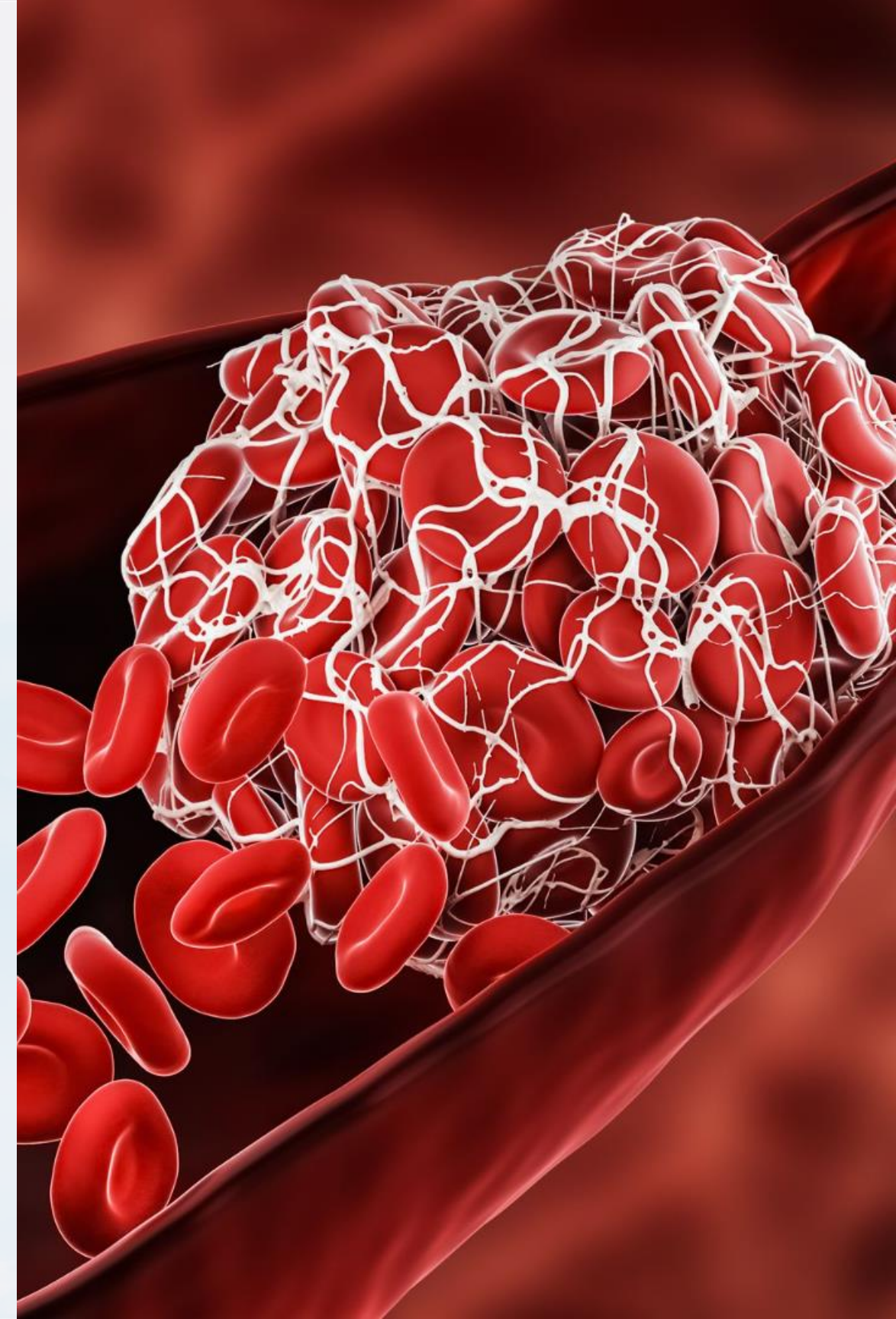


Decoding DVT: An Investigation of Contributing Factors with Machine Learning.

Understanding Deep Vein Thrombosis

Deep vein thrombosis (DVT) is a condition where blood clots form in the deep in the deep veins, typically in the legs. It can be serious if a clot breaks free and breaks free and travels to the lungs, causing a pulmonary embolism.





Reason for choosing this problem

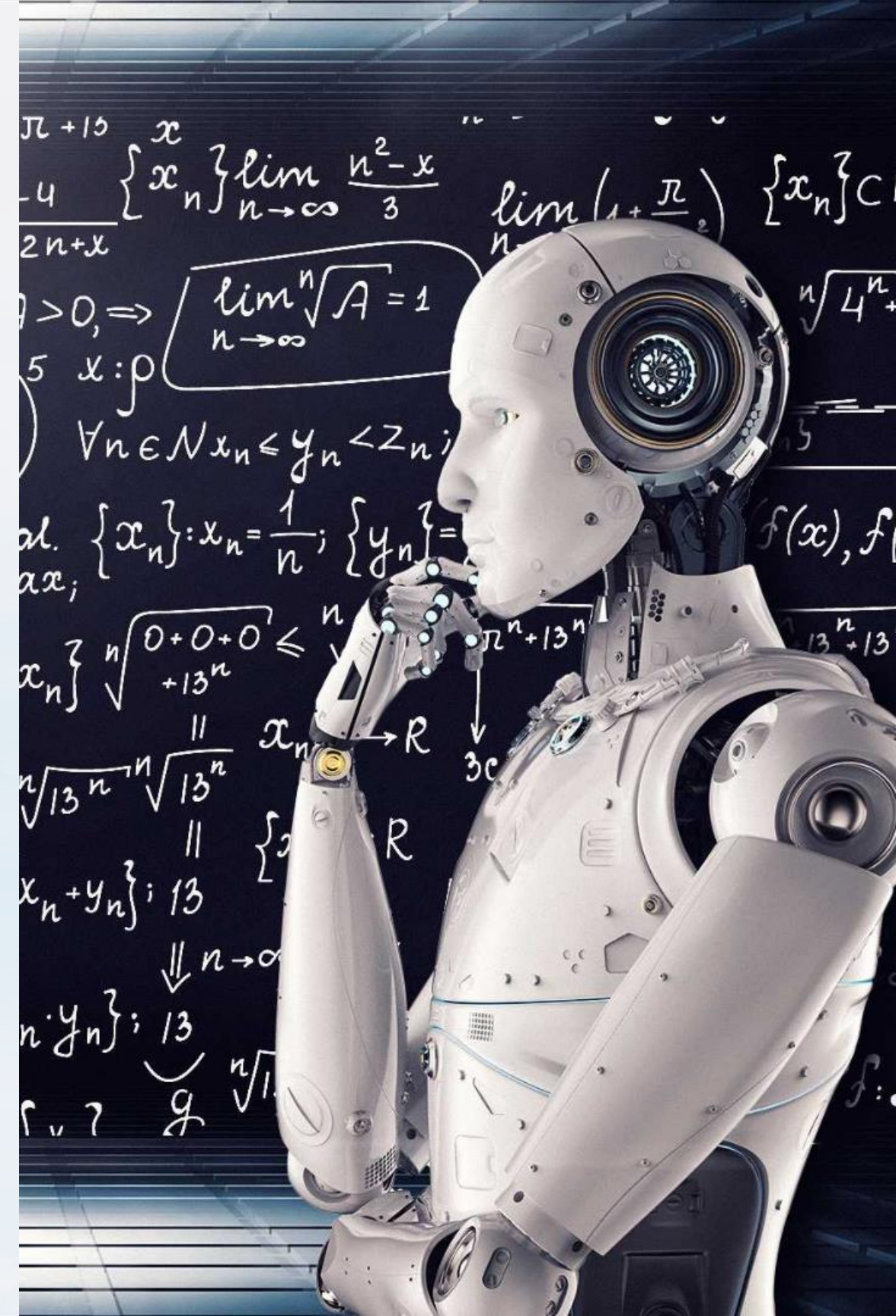
One member of our group was diagnosed with deep vein thrombosis (DVT) later than they should have been. We learned that DVT is often misdiagnosed because its symptoms can be similar to other conditions. We want to find a better way to diagnose this disease so that others don't have to go through the same experience.

The Goal of Our Project

DVT is a serious condition that can be fatal if not diagnosed and treated early. Our goal is to identify factors that identify factors that increase the risk of DVT so that we can develop better prevention and treatment strategies. treatment strategies.

Machine Learning

Machine learning is a field of study that uses statistical models and algorithms to enable computers to learn from data without being explicitly programmed.





Unlocking Insights with UK Biobank Data

The UK Biobank is a large-scale biomedical database that is held in the UK. The database contains detailed medical information about more than 500,000 individuals.

Overview of the Dataset

500,000+

Number of individuals in the dataset.

4

Number of time points in the dataset.

Sparse Data

Many of the features are missing or irrelevant, irrelevant, making feature extraction a challenge. challenge.

Mixed Data Types

The database includes both categorical and and continuous data, which requires special special processing to use as input for machine machine learning models.



Technical Difficulties

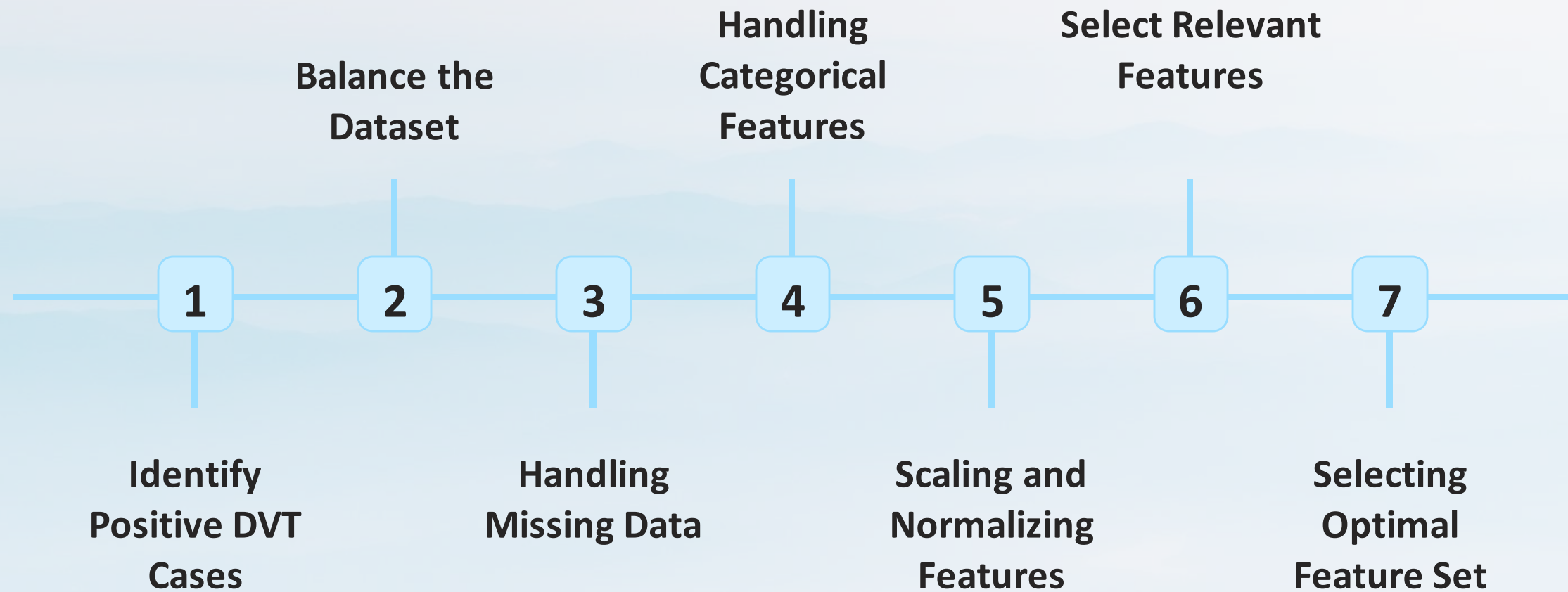
Problems:

- Remote data
- Vpn with Internet block
- Automatic disconnect

Solutions :

- Logging
- Virtual machine
- Extensions

Research Process



Preparing The Dataset

Identify Positive DVT Cases :

- Filtering the UK Biobank data to find all patients who have had at least one diagnosed instance of DVT.

Balance the Dataset :

- We need to balance the dataset by randomly selecting an equal number of patients who have never had DVT.

Handling the Data

Handling Missing Data :

- Many features are missing a large percentage of their values, We developed strategies to handle missing handle missing data.

Handling Categorical Features :

- We used the OneHot techniques, OneHot encoding transforms categorical features into binary vectors, binary vectors, where each category is represented by a 1 (presence) or 0 (absence), enabling their use enabling their use in machine learning models.

Scaling and Normalizing Features :

- We must standardize the scale of features with different units and ranges to make them comparable and comparable and enhance the performance of machine learning models.

Feature Selection

Select Relevant Features :

We need to identify which features are relevant or irrelevant to DVT prediction. This requires careful analysis and domain knowledge.

Selecting Optimal Feature Set :

Not all features are equally important in predicting DVT. We must identify the most relevant subset of features to achieve the best performance.

Analysis

We employed two primary evaluation metrics for our models:

1. **F1 Score:** A balanced measure of precision and recall, essential for imbalanced medical datasets.
2. **Confusion Matrix:** This visual tool summarizes model performance, It performance, It aids in error analysis and model refinement.

Additionally, we utilized several other evaluation metrics to comprehensively assess our models' performance.



Conclusion

Impact

Our work has a significant impact on understanding DVT development from various biomedical perspectives, aiding early diagnosis, treatment, and prevention.

Application

Our machine learning model can be used to predict the risk of DVT, providing recommendations for better prevention and management of the disease.

Future Work

In the future, we can incorporate additional datasets and sources to further improve the accuracy and robustness of the machine learning model.