UNIVERSITI MALAYA

FACULTY OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY

SEMESTER 1 2024/2025

WQD7005 DATA MINING (OCC 1)

PROJECT

LECTURER: PROF DR TEH YING WAH

NAME: YEE SEE MARN

MATRIC NUMBER: 23102510

Comprehensive AI-Assisted Final Report

Dataset Simulation and Feature Engineering

The dataset in this project is generated by using the GenAI model (GPT-4o) through several designs of prompts that contain 15,000 synthetic patient records with 10 numerical features, 4 categorical features, clinical notes and 'deteriorated' as the target variables. The features are designed according to the references (Medscape, 2024) and (Riley, 2025). Figure 1 shows the percentage of missing values of each column through *y-data profiling*. Figure 2 shows the steps of feature engineering on the columns with missing values accordingly, while Figure 3 shows the extraction of meaningful features from the clinical notes.

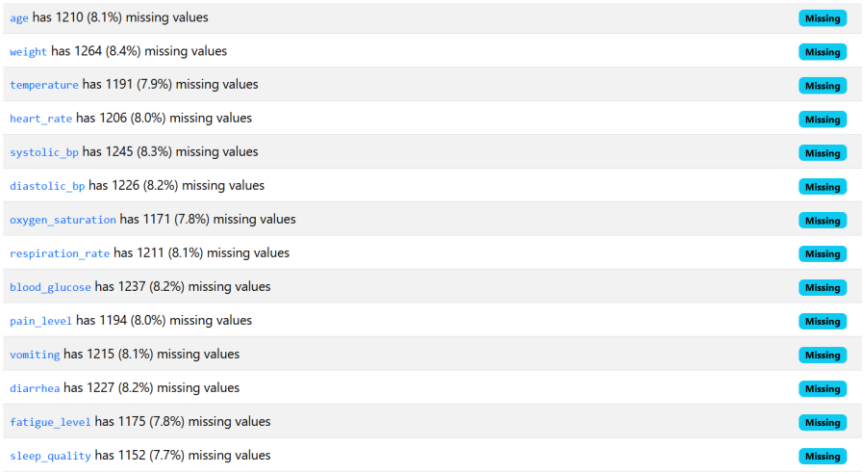


Figure 1: Percentage of missing values of each column

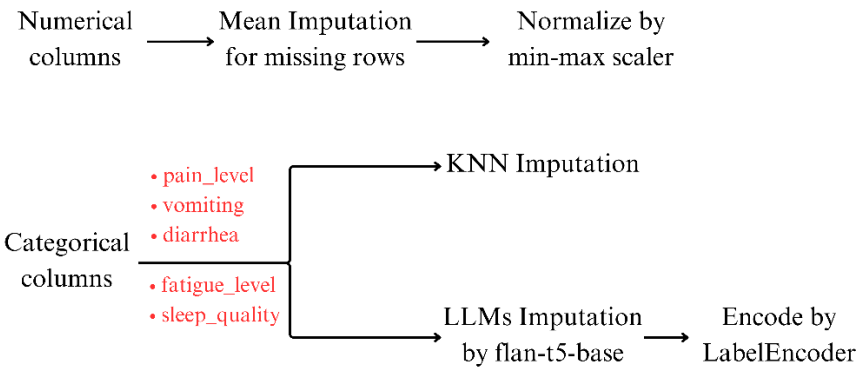


Figure 2: Feature Engineering for Numerical and Categorical Columns

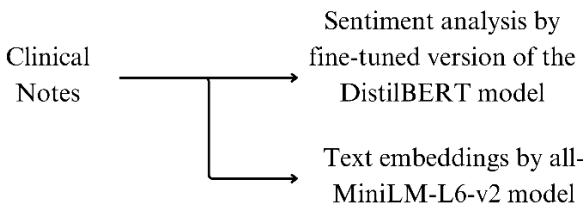


Figure 3: Applying LLMs to Clinical Notes

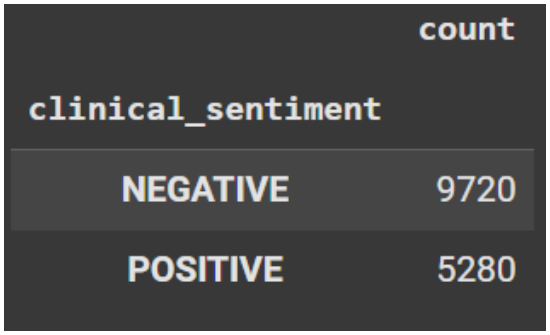
Predictive Model Development

The dataset is split to 80% as training set while 20% as testing set. Random Forest, XGBoost, Neural Network and Multi-Modal Transformer are chosen to construct the predictive model. Random Forest is a machine learning algorithm that combines the output of multiple decision trees to make predictions. XGBoost is the use of gradient boosting to build predictive models. Neural Networks are a machine learning model that inspired by the structure of human brains that consists of connected units or nodes. In this project, sequential model through *tensor flow keras* is used. An advanced transformer model of Multi-Modal is a type of neural network architecture that can integrate and understand data from multiple sources.

Some important parameters used for each model:

- Random Forest – number of estimators
- XGBoost – evaluation metrics
- Sequential based Neural Network – Number of neurons, activation function, optimizer, loss function
- Multi-Modal - Number of neurons, activation function, optimizer, loss function, multi-head attention

Sentiment analysis of clinical notes are also done by utilizing the *distilbert-base-uncased-finetuned-sst-2-english* model and Figure 4 shows the value counts of positive and negative sentiment.



clinical_sentiment	count
NEGATIVE	9720
POSITIVE	5280

Figure 4: Value Counts

Model Evaluation and Interpretation

Table 1 show the accuracy, precision, recall, F1 score, and confusion matrix of different machine learning models.

Table 1: Accuracy, Precision, Recall, F1 Score, and Confusion Matrix of the Models

	Accuracy	Precision	Recall	F1-Score	Confusion Matrix
Random Forest	0.8613	0.9894	0.3100	0.4721	[[2398 2] [414 186]]
XGBoost	0.8490	0.7348	0.3833	0.5038	[[2317 83] [370 230]]
Neural Network	0.8430	0.9108	0.2383	0.3378	[[2386 14] [457 143]]
Multi-Modal Transformer	0.8533	0.8883	0.3050	0.4541	[[2377 23] [417 183]]

In general, all the models achieve an accuracy of 80% and above. However, it is also critical to increase the value of recall and f1-score. This issue may be due to the imbalanced dataset. Thus, we can consider techniques like oversampling the minority class or undersampling the majority class for future improvements to enhance the recall and f1-score.

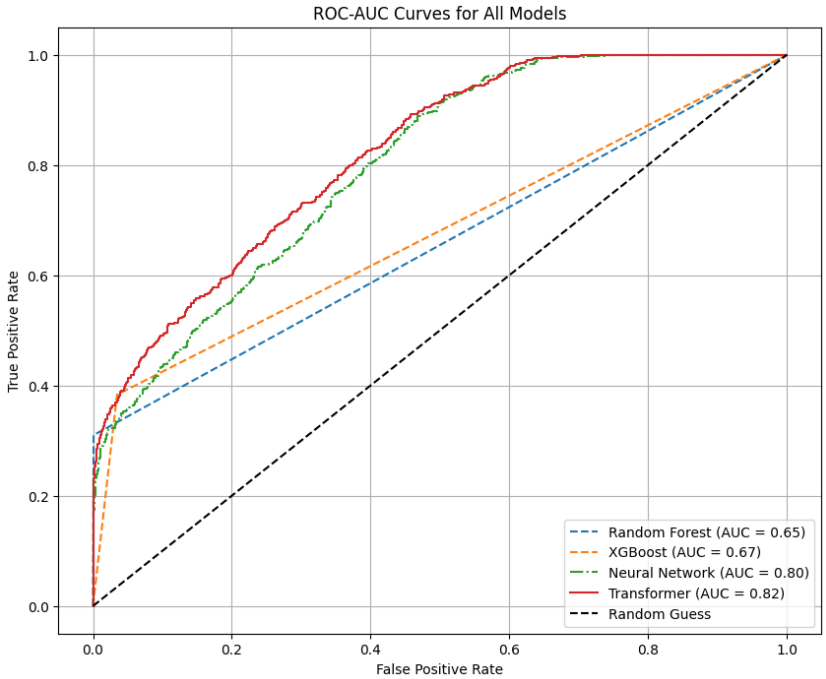


Figure 5: ROC Curve for All Models

Figure 5 shows the graph of Receiver Operating Characteristic (ROC) curve for all the machine learning models and Multi-Modal Transformer's curve is the closest to the top-left corner, indicating the best trade-off between true positive rate and false positive rate.

Next, we calculate the top three features and the SHAP value that contribute to each of the predictive models. Only one instance prediction is used to calculate the SHAP value, and the summary of the interpretation is generated by leveraging the *gemini-2.0-flash* model.

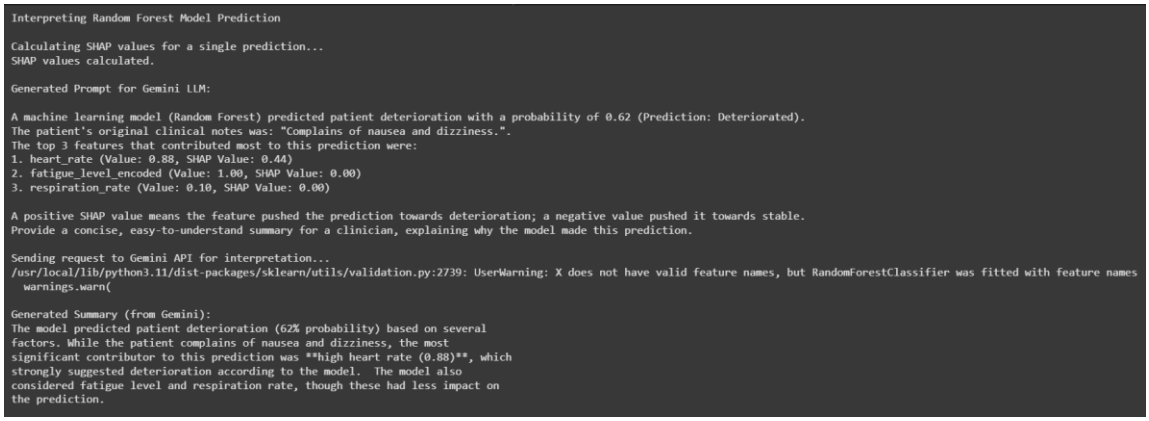


Figure 6: Interpretation of Random Forest

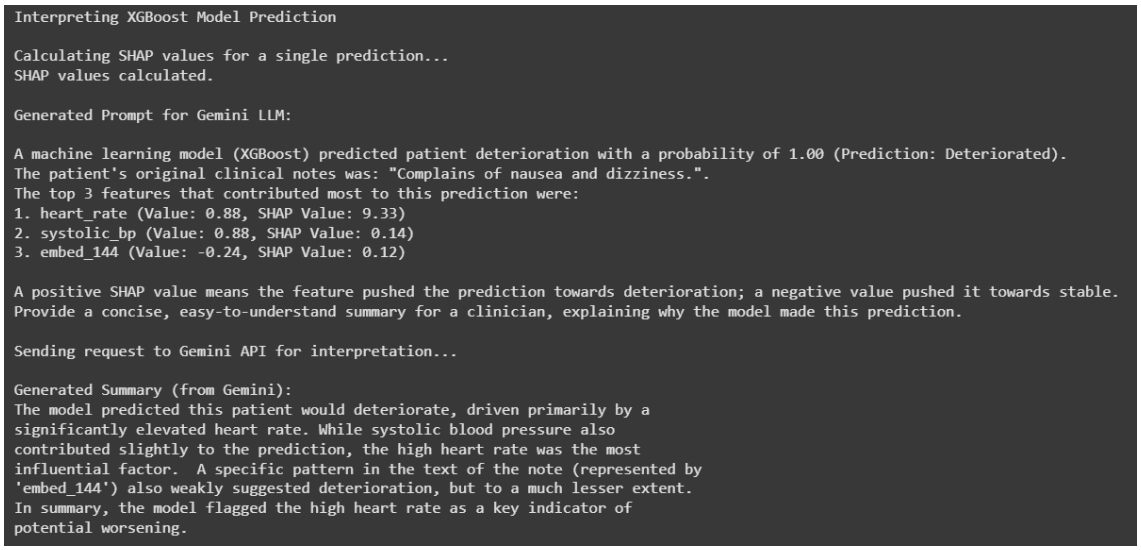


Figure 7: Interpretation of XGBoost

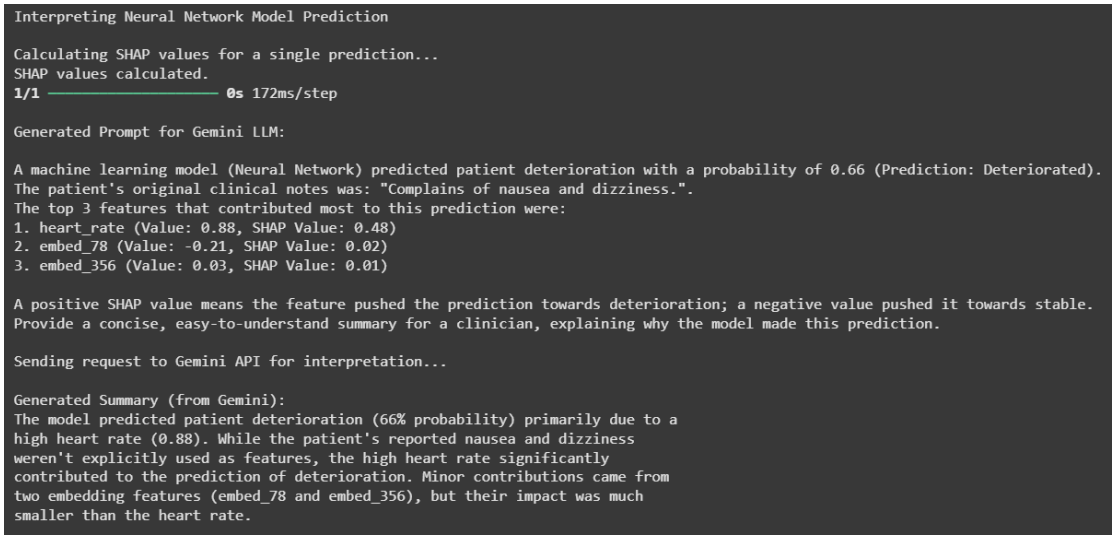


Figure 8: Interpretation of Neural Network

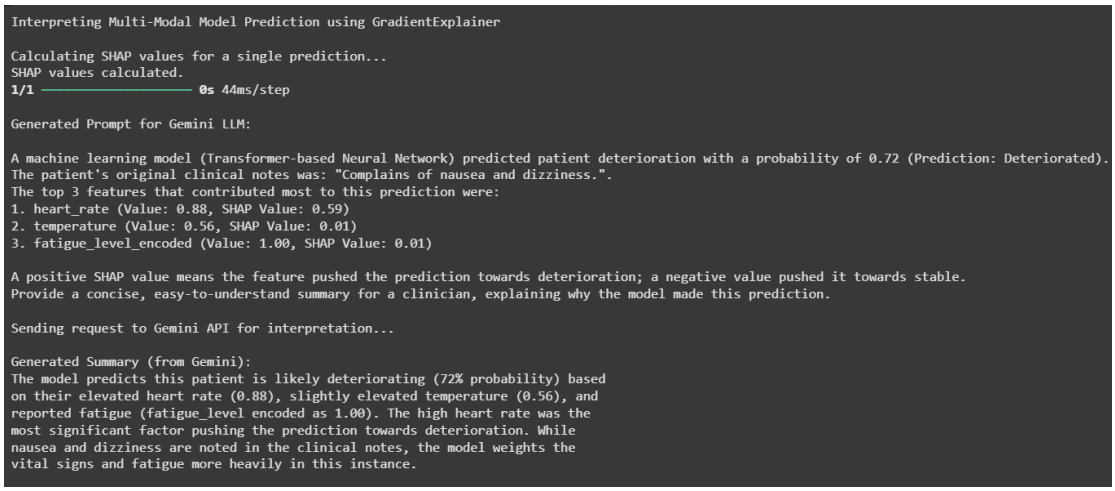


Figure 9: Interpretation of Multi-Modal Transformer

Table 2: Comparison of features importance for each model for one instance

Model	Features	SHAP Value
Random Forest	• Hear_rate • Fatigue_level_encoded • Respiration_rate	• 0.44 • 0.00 • 0.00
XGBoost	• Heart_rate • Systolic_bp • Embed_144	• 9.33 • 0.14 • 0.12
Neural Network	• Heart_rate • Embed_78 • Embed_356	• 0.48 • 0.02 • 0.01

Multi-Modal Transformer	• Heart_rate • Temperature • Fatigue_level_encoded	• 0.59 • 0.01 • 0.01

Conclusion and Recommendations

In conclusion, XGBoost is the best model to be chosen for this stage, as its balance between the evaluation metrics and heart rate appears as the most important feature in all the predictive models. However, we recommend fine-tuning each model to enhance their performance. Balancing of the datasets by GenAI should also be considered to improve the power of the predictive model. SHAP value across the dataset should be calculated instead of just a single instance in future works.

References

Medscape. (2024, February 16). *Normal Vital Signs: Normal Vital Signs, Normal Heart Rate, Normal Respiratory Rate*. Medscape.com. https://emedicine.medscape.com/article/2172054-overview#showall

Riley, L. (2025). *Mean Fasting Blood Glucose*. World Health Organization. https://www.who.int/data/gho/indicator-metadata-registry/imr-details/2380