



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.

<code>age</code> has 1210 (8.1%) missing values	Missing
<code>weight</code> has 1264 (8.4%) missing values	Missing
<code>temperature</code> has 1191 (7.9%) missing values	Missing
<code>heart_rate</code> has 1206 (8.0%) missing values	Missing
<code>systolic_bp</code> has 1245 (8.3%) missing values	Missing
<code>diastolic_bp</code> has 1226 (8.2%) missing values	Missing
<code>oxygen_saturation</code> has 1171 (7.8%) missing values	Missing
<code>respiration_rate</code> has 1211 (8.1%) missing values	Missing
<code>blood_glucose</code> has 1237 (8.2%) missing values	Missing
<code>pain_level</code> has 1194 (8.0%) missing values	Missing
<code>vomiting</code> has 1215 (8.1%) missing values	Missing
<code>diarrhea</code> has 1227 (8.2%) missing values	Missing
<code>fatigue_level</code> has 1175 (7.8%) missing values	Missing
<code>sleep_quality</code> has 1152 (7.7%) missing values	Missing

Figure 1: Percentage of missing values of each column

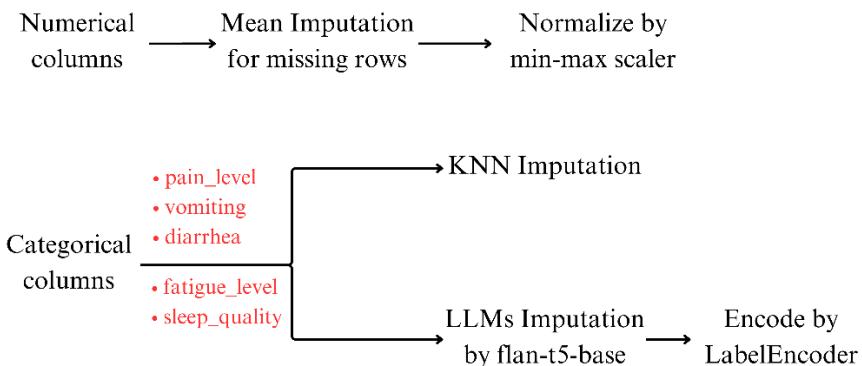


Figure 2: Feature Engineering for Numerical and Categorical Columns

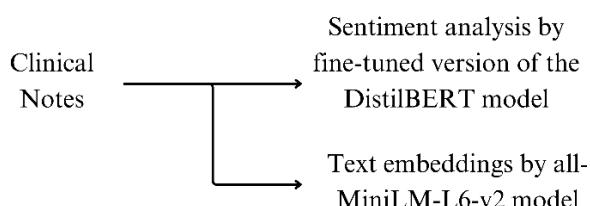


Figure 3: Applying LLMs to Clinical Notes

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.

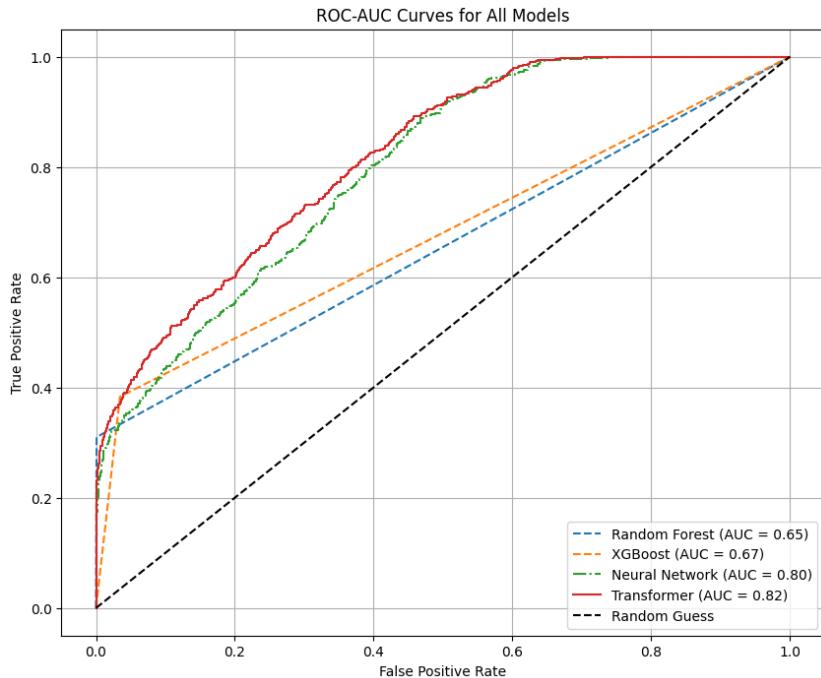


Figure 5: ROC Curve for All Models

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.

```
Interpreting Random Forest Model Prediction
Calculating SHAP values for a single prediction...
SHAP values calculated.

Generated Prompt for Gemini LLM:
A machine learning model (Random Forest) predicted patient deterioration with a probability of 0.62 (Prediction: Deteriorated).
The patient's original clinical notes was: "Complains of nausea and dizziness.".
The top 3 features that contributed most to this prediction were:
1. heart_rate (Value: 0.88, SHAP Value: 0.44)
2. fatigue_level_encoded (Value: 1.00, SHAP Value: 0.00)
3. respiration_rate (Value: 0.10, SHAP Value: 0.00)

A positive SHAP value means the feature pushed the prediction towards deterioration; a negative value pushed it towards stable.
Provide a concise, easy-to-understand summary for a clinician, explaining why the model made this prediction.

Sending request to Gemini API for interpretation...
/usr/local/lib/python3.11/dist-packages/scikitlearn/utils/validation.py:2739: UserWarning: X does not have valid feature names, but RandomForestClassifier was fitted with feature names
warnings.warn(
Generated Summary (from Gemini):
The model predicted patient deterioration (62% probability) based on several
factors. While the patient complains of nausea and dizziness, the most
significant contributor to this prediction was **high heart rate (0.88)**, which
strongly suggested deterioration according to the model. The model also
considered fatigue level and respiration rate, though these had less impact on
the prediction.
```

Figure 6: Interpretation of Random Forest

```
Interpreting XGBoost Model Prediction
Calculating SHAP values for a single prediction...
SHAP values calculated.

Generated Prompt for Gemini LLM:
A machine learning model (XGBoost) predicted patient deterioration with a probability of 1.00 (Prediction: Deteriorated).
The patient's original clinical notes was: "Complains of nausea and dizziness.".
The top 3 features that contributed most to this prediction were:
1. heart_rate (Value: 0.88, SHAP Value: 9.33)
2. systolic_bp (Value: 0.88, SHAP Value: 0.14)
3. embed_144 (Value: -0.24, SHAP Value: 0.12)

A positive SHAP value means the feature pushed the prediction towards deterioration; a negative value pushed it towards stable.
Provide a concise, easy-to-understand summary for a clinician, explaining why the model made this prediction.

Sending request to Gemini API for interpretation...

Generated Summary (from Gemini):
The model predicted this patient would deteriorate, driven primarily by a
significantly elevated heart rate. While systolic blood pressure also
contributed slightly to the prediction, the high heart rate was the most
influential factor. A specific pattern in the text of the note (represented by
'embed_144') also weakly suggested deterioration, but to a much lesser extent.
In summary, the model flagged the high heart rate as a key indicator of
potential worsening.
```

Figure 7: Interpretation of XGBoost

```

Interpreting Neural Network Model Prediction

Calculating SHAP values for a single prediction...
SHAP values calculated.
1/1 ━━━━━━ 0s 172ms/step

Generated Prompt for Gemini LLM:

A machine learning model (Neural Network) predicted patient deterioration with a probability of 0.66 (Prediction: Deteriorated). The patient's original clinical notes was: "Complains of nausea and dizziness.". The top 3 features that contributed most to this prediction were:
1. heart_rate (Value: 0.88, SHAP Value: 0.48)
2. embed_78 (Value: -0.21, SHAP Value: 0.02)
3. embed_356 (Value: 0.03, SHAP Value: 0.01)

A positive SHAP value means the feature pushed the prediction towards deterioration; a negative value pushed it towards stable. Provide a concise, easy-to-understand summary for a clinician, explaining why the model made this prediction.

Sending request to Gemini API for interpretation...

Generated Summary (from Gemini):
The model predicted patient deterioration (66% probability) primarily due to a high heart rate (0.88). While the patient's reported nausea and dizziness weren't explicitly used as features, the high heart rate significantly contributed to the prediction of deterioration. Minor contributions came from two embedding features (embed_78 and embed_356), but their impact was much smaller than the heart rate.

```

Figure 8: Interpretation of Neural Network

```

Interpreting Multi-Modal Model Prediction using GradientExplainer

Calculating SHAP values for a single prediction...
SHAP values calculated.
1/1 ━━━━━━ 0s 44ms/step

Generated Prompt for Gemini LLM:

A machine learning model (Transformer-based Neural Network) predicted patient deterioration with a probability of 0.72 (Prediction: Deteriorated). The patient's original clinical notes was: "Complains of nausea and dizziness.". The top 3 features that contributed most to this prediction were:
1. heart_rate (Value: 0.88, SHAP Value: 0.59)
2. temperature (Value: 0.56, SHAP Value: 0.01)
3. fatigue_level_encoded (Value: 1.00, SHAP Value: 0.01)

A positive SHAP value means the feature pushed the prediction towards deterioration; a negative value pushed it towards stable. Provide a concise, easy-to-understand summary for a clinician, explaining why the model made this prediction.

Sending request to Gemini API for interpretation...

Generated Summary (from Gemini):
The model predicts this patient is likely deteriorating (72% probability) based on their elevated heart rate (0.88), slightly elevated temperature (0.56), and reported fatigue (fatigue_level encoded as 1.00). The high heart rate was the most significant factor pushing the prediction towards deterioration. While nausea and dizziness are noted in the clinical notes, the model weights the vital signs and fatigue more heavily in this instance.

```

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	<ul style="list-style-type: none"> • Hear_rate • Fatigue_level_encoded • Respiration_rate 	<ul style="list-style-type: none"> • 0.44 • 0.00 • 0.00
XGBoost	<ul style="list-style-type: none"> • Heart_rate • Systolic_bp • Embed_144 	<ul style="list-style-type: none"> • 9.33 • 0.14 • 0.12
Neural Network	<ul style="list-style-type: none"> • Heart_rate • Embed_78 • Embed_356 	<ul style="list-style-type: none"> • 0.48 • 0.02 • 0.01

Multi-Modal Transformer	<ul style="list-style-type: none"> • Heart_rate • Temperature • Fatigue_level_encoded 	<ul style="list-style-type: none"> • 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>