Model Performance Comparison: Predicting Concentration

This report compares the performance of two models developed to predict patient 'concentration' levels, now framed as a **5-class classification problem**. The original 'concentration' ratings (-2, -1, 0, 1, 2) are assumed to correspond to classes 0, 1, 2, 3, 4, respectively.

1. **Random Forest Classifier (Baseline)**: Utilized only tabular data.
2. **CLSTM (Convolutional LSTM)**: A more complex deep learning model that combines tabular data with textual features extracted from 'doctor_notes'.

The goal is to determine which model performs better and to understand the implications of their results in this classification context.

Performance Metrics Overview

Metric	Random Forest Classifier	CLSTM (Tabular + Text)	Interpretation
Accuracy	0.2003 (20.03%)	0.1992 (19.92%)	Percentage of correct predictions. Both are very low.
F1 Score (Weighted)	0.2002	0.1712	Harmonic mean of precision & recall. Both are very low.
Test Loss (CLSTM)	N/A	1.6098	Likely Categorical Crossentropy.

Note: For a 5-class problem with balanced classes (as suggested by the support in the classification reports), an accuracy of around 20% (1/5) is expected from random guessing.

Detailed Interpretation

1. Random Forest Classifier (Baseline)

- **Accuracy: 0.2003**
  - The model correctly predicts the concentration class for approximately 20.03% of the patients in the test set. This is essentially at the level of random chance for a 5-class problem.
- **F1 Score (Weighted): 0.2002**
  - The weighted F1-score, which accounts for class imbalance (though classes

seem fairly balanced here), is also very low, indicating poor performance across precision and recall.

- Classification Report Breakdown:
Class (Assumed Original)	Precision	Recall	F1-Score	Support
0 (-2)	0.19	0.20	0.20	799
1 (-1)	0.20	0.20	0.20	803
2 (0)	0.20	0.20	0.20	800
3 (1)	0.23	0.22	0.22	792
4 (2)	0.19	0.17	0.18	806
  - **Interpretation**: The Random Forest Classifier performs very similarly across all classes, with precision, recall, and F1-scores hovering around 0.20 for each. This uniformity at a low performance level further reinforces the idea that the model is not learning any meaningful patterns to distinguish between the concentration classes. Class 3 (assumed rating '1') shows marginally better precision and F1-score, but the difference is minimal.

2. CLSTM (Convolutional LSTM - Tabular + Text) - Classification

- **Test Accuracy: 0.1992**
  - The CLSTM model's accuracy (19.92%) is also at the random chance level, and marginally lower than the Random Forest.
- **Test F1 Score (Weighted): 0.1712**
  - The weighted F1-score for the CLSTM is notably lower than the Random Forest's, suggesting even poorer combined precision and recall performance, especially when considering class-specific metrics.
- **Test Loss: 1.6098**
  - This is likely the categorical cross-entropy loss, a standard loss function for multi-class classification.
- Classification Report Breakdown (CLSTM):
Class (Assumed Original)	Precision	Recall	F1-Score	Support
0 (-2)	0.24	0.09	0.13	799
1 (-1)	0.19	0.25	0.22	803
2 (0)	0.21	0.28	0.24	800
3 (1)	0.16	0.01	0.02	792
4 (2)	0.19	0.37	0.25	806
  - **Interpretation**: The CLSTM model exhibits more varied performance across classes compared to the Random Forest, but it's generally poor:
    - **Class 3 (assumed rating '1')**: Performance is extremely poor, with a recall

of only 0.01 and an F1-score of 0.02. The model almost completely fails to identify this class.
    - **Class 0 (assumed rating '-2')**: Shows relatively higher precision (0.24) but very low recall (0.09).
    - **Class 4 (assumed rating '2')**: Has the highest recall (0.37), meaning it identifies this class more often when it occurs, but its precision (0.19) is low, indicating many false positives for this class.
    - The model seems to be biased towards predicting some classes more than others, but not accurately. The low weighted F1-score reflects this imbalanced and poor per-class performance.

Comparative Analysis

- **Overall Performance (Accuracy & F1-Score)**:
    - Both models perform at a level consistent with random guessing for a 5-class problem (around 20% accuracy).
    - The **Random Forest Classifier** has slightly better overall accuracy (20.03% vs. 19.92%) and a better weighted F1-score (0.2002 vs. 0.1712) than the CLSTM.
    - This indicates that, in this classification setup, the simpler Random Forest model using only tabular data outperformed the more complex CLSTM model that also incorporated text data.
- **Consistency vs. Variability**:
    - The Random Forest showed consistently poor performance across all classes.
    - The CLSTM showed more erratic performance, doing extremely poorly on some classes (e.g., class 3 recall of 0.01) while having slightly higher recall on others (e.g., class 4 recall of 0.37), but still with low precision.
- **Impact of Text Data (CLSTM)**:
    - The inclusion of text data and the more complex CLSTM architecture resulted in *worse* performance (lower accuracy and significantly lower weighted F1-score) compared to the Random Forest using only tabular data. This suggests the text features, as processed, might have introduced more noise than signal for this classification setup, or the model struggled to effectively learn from the combined features for this specific task.

Conclusion and Recommendations

1. **Overall Performance**: Both models are not effective at predicting *concentration* when framed as a 5-class classification problem. Their performance is at or near random chance.
2. **Model Choice**: In this classification scenario, the **Random Forest Classifier**

performed slightly better than the CLSTM in terms of overall accuracy and weighted F1-score, despite being a simpler model.

3. **Predictive Power of Features**: The results strongly suggest that the available features (both tabular and the *doctor_notes* text) do not contain sufficient discriminatory information to reliably classify patients into one of the five concentration levels.
4. **Potential Next Steps**:
    - **Confirm Class Mapping**: Ensure the mapping from original ratings (-2 to 2) to classes (0 to 4) is correct and consistently applied.
    - **Simplify the Problem**:
        - Consider reducing the number of classes. For example, could ***concentration*** be meaningfully grouped into fewer categories (e.g., Low [-2, -1], Neutral [0], High [1, 2])? This might make the classification task easier if finer distinctions are too noisy.
        - Revisit if regression is indeed a more suitable framework, despite its initial poor R^2 values that it produced. The fundamental issue seems to be the lack of predictive signal.
    - **Feature Engineering**: This remains crucial. The current features are not working.
        - *Text Data*: For CLSTM, explore more advanced NLP techniques: pre-trained embeddings (Word2Vec, GloVe, BERT), attention mechanisms, or different text preprocessing strategies. The current 822 unique tokens might be too small or too generic.
        - *Tabular Data*: Look for more relevant tabular features or create interaction terms.
    - **Data Augmentation/Collection**: The dataset might be insufficient in size or quality to capture the complex nuances of *concentration*.
    - **Domain Expertise**: This is paramount. Consult with medical or psychological experts to understand:
        - What are the truly reliable indicators of different concentration levels?
        - Are the *doctor_notes* actually capturing information relevant to concentration, or are they too general? Could more structured note-taking or specific questionnaires provide better data?
        - Is the 5-point rating scale for *concentration* robust and consistently applied by raters?

The core challenge seems to lie in the inherent difficulty of predicting *concentration* from the available data, regardless of whether it's treated as a regression or classification task.