The Academic College of Tel Aviv-Yaffo

THE SCHOOL OF COMPUTER SCIENCE

Harmony in Diagnosis: Exploring Consensus and Variability in Clinical Judgement

September 2024

Thesis submitted in partial fulfilment of the requirements for the M.Sc. degree in the School of Computer Science of the Academic College of Tel Aviv-Yaffo

By

Sanad Satel

The research work for the thesis has been carried out under the supervision of

Dr. Sarel Cohen

Contents

1 abstract

This research explores the variability in decision-making among clinical psychologists by applying machine learning techniques to reverse-engineer their judgments. Using a dataset of 861 patients evaluated by 29 judges (both experienced psychologists and trainees), we trained decision tree and linear regression models for each judge to capture their decision patterns. We employed methods such as Jaccard similarity and pairwise Mean Squared Error (MSE) to quantify the distance between judges' models, providing a clear metric of variability. Additionally, we applied SMOTE for data augmentation to enhance model training and improve robustness in comparisons. To gain deeper insights into the decision-making process, we used Large Language Models (LLMs) to explain and compare individual and pairwise models, highlighting key differences in clinical judgment. Our findings reveal significant variability in how judges weigh psychological traits, offering valuable insights into the potential for improving consistency and understanding judgment discrepancies in clinical practice. You can also find the related code in the following GitHub Repository

The research results conducted during the course of the work on this thesis were submitted to ComplexNetworks'24 conference.

2 Introduction

Clinical judgment plays a critical role in diagnosing and treating mental health disorders, yet variability among expert judgments remains a significant concern. Even among experienced clinical psychologists, decisions regarding the same patient can differ, leading to inconsistent diagnoses and potentially divergent treatment paths. Understanding the factors that contribute to this variability is crucial for improving the consistency and reliability of mental health diagnoses.

One of the most influential datasets for studying clinical judgment variability is Meehl's dataset Meehl, which provides a source of patient evaluations by multiple clinical judges. This dataset has been used extensively in psychological research to explore how different judges assess the same psychological profiles, but much remains unknown about the underlying decision-making processes and the extent of variability between experts.

Recent advances in machine learning provide new opportunities to analyze and model clinical judgment. By training machine learning models on the decisions made by different judges, it becomes possible to quantify the similarities and differences between their decision-making processes. Furthermore, the explainability of these models can offer valuable insights into the factors driving variability in clinical judgment.

In this research, we aim to quantify and analyze judgment variability among clinical psychologists using machine learning techniques. By applying decision tree and linear regression models to predict clinical judgments based on patient characteristics, we measure the distance between different judges' decision-making patterns using metrics such as Jaccard similarity and Mean Squared Error (MSE). Additionally, we leverage Large Language Models (LLMs) to explain and compare individual and pairwise decision models, providing a deeper understanding of the key differences between judges.

Our work contributes to the growing field of machine learning applications in mental health by offering a few approaches to modeling judgment variability. By examining the underlying patterns in clinical judgments, we aim to shed light on the factors that contribute to variability and propose ways to improve consistency in clinical decision-making.

This thesis is organized as follows. In Section 3 we discuss related work. Section 4 describes our dataset. Section 5 outlines the experiments we want to do in the research Section 6 we show the results and deeply describe the process. Finally, in Section 7, we conclude and suggest future research directions.

3 Related Work

The study of clinical judgment variability has a long history, particularly in the context of mental health diagnoses. A significant portion of the early work in this area was pioneered by Paul Meehl, whose research on the application of statistical models in clinical settings highlighted the limitations of subjective expert judgment. Meehl argued that statistical models often outperform human

clinicians when it comes to making consistent and reliable diagnoses Meehl. His dataset, used extensively in our study, provides an empirical basis for exploring patterns in clinical judgments and the factors that contribute to variability.

One prominent study that expands upon Meehl's work is by Yoav Ganzach, who explored the use of non-linear models in clinical judgment. Ganzach demonstrated that non-linear models, such as decision trees, are better suited to capturing the complex decision-making processes of clinicians compared to linear models Ganzach [2001]. His research, which also utilizes Meehl's dataset, provides key insights into the types of models that are most effective for modeling clinical decisions, supporting our choice of decision tree models in this study. Furthermore, Ganzach's work on non-linearities helps to justify our approach to comparing judges by focusing on the structure and variability in their decision-making processes.

Furthermore, the role of machine learning in clinical decision-making is increasingly being recognized, particularly for its potential to augment expert judgment. Recent research by Adlung et al. Adlung et al. [2022] discusses how machine learning can help to process complex, heterogeneous data to support clinical decisions. Machine learning models are well-suited to integrate vast amounts of patient data and improve decision consistency, which aligns with our goal of using these models to quantify judgment variability. However, the application of these techniques to high-level decision-making still requires further validation, highlighting the challenges of integrating machine learning in clinical practice.

More recently, the explainability of machine learning models has become a focus of research, particularly in clinical settings where decisions must be interpretable for practitioners. Miller et al. Miller et al. [2017] reviewed various techniques for model explainability, underscoring the need for transparency in models used in healthcare. Our use of Large Language Models (LLMs) to explain the differences between judges' decision-making processes aligns with this growing emphasis on model interpretability. By incorporating LLM-based explanations, we aim to provide a clearer understanding of the judgment variability among clinicians, offering insights that are not only accurate but also interpretable.

Our work builds upon these contributions by applying machine learning techniques, particularly decision trees and linear regression models, to quantify judgment variability among clinical psychologists. We extend the existing literature by incorporating data augmentation techniques, such as SMOTE, to ensure robust model training and by utilizing Large Language Models (LLMs) to provide textual explanations of the models' decision-making patterns.

4 Dataset

The dataset used in this research is derived from Paul Meehl's extensive studies on clinical judgment. It contains evaluations of 861 patients diagnosed as either neurotic or psychotic based on the Minnesota Multiphasic Personality Inventory

(MMPI), a widely used psychological assessment tool - heir scores on eight clinical scales and three validity scales of the MMPI, where the evaluations were on a scale between 0 (psychotis) to 11 (neurosis).

Additionally, the dataset includes judgments from 29 clinical experts, comprising 13 professional clinical psychologists and 16 clinical psychology trainees. Each judge's evaluation corresponds to one of the 29 target columns in the dataset, representing their assessment of each patient.

The dataset consists of two key components:

- **MEELMMPI**: This dataset contains the features X for each of the 861 patients, capturing the 11 psychological traits based on the MMPI assessment.
- **MEELJUD**: This dataset contains the target variables y, representing the judgments made by the 29 judges. Each column in this dataset corresponds to the judgment of a single judge, where each row represents the evaluation of a given patient.

The characteristics are lie, eccentricity, defensiveness, hypochondriasis, depression, hysteria, psychopathic deviate, paranoia, psychasthenia, schizophrenia, and hypomania, total of 11 characteristics. Each characteristic plays a crucial role in understanding and assessing an individual's personality. A brief overview can be seen in Table 4.

The goal of this study is to model the variability in judgment across the different judges by analyzing the differences in their decision-making processes. We utilize both decision tree and linear regression models to capture these differences, and we employ various metrics, such as Mean Squared Error (MSE) to quantify the variability between judges.

Characteristic	Meaning
Lie	Assesses the tendency to present oneself in a favorable light, detecting potential deception or social desirability bias.
Eccentricity	Measures unconventional behavior or thinking patterns, offering insights into an individual's uniqueness.
Defensiveness	Indicates the inclination to avoid revealing personal vulnerabilities, reflecting defensive mechanisms.
Hypochondriasis	Assesses the tendency to be preoccupied with physical health and exhibit somatic concerns.
Depression	Measures the presence and intensity of depressive symptoms, aiding in mood disorder assessment.
Hysteria	Evaluates the expression of emotional distress, providing insights into potential conversion disorders.
Psychopathic Deviate	Assesses tendencies toward social deviance, unconventional behavior, and disregard for social norms.
Paranoia	Measures distrust and suspicion of others, helping identify paranoid tendencies.
Psychasthenia	Indicates the presence of obsessive-compulsive symptoms, measuring anxiety-related traits.
Schizophrenia	Assesses characteristics associated with schizophrenia spectrum disorders, aiding in psychosis evaluation.
Hypomania	Measures elevated mood, impulsivity, and energy levels, providing insights into manic tendencies.

Table 1: The 11 Charactaristics of the MMPI-2 Personality Disorder Spectra Scales, as defined in Mulay et al. [2019].

5 Methodology

5.1 Experimental Overview and Key Findings

To explore the decision-making patterns among clinical judges, we employed various machine learning techniques. This section details the approaches used, including model training, cross-judge predictions, and the assessment of feature importance. By applying these methods, we aimed to quantify the similarities and differences in judgments and identify key factors influencing decision accuracy.

5.2 Estimate the Distance between Judges

We employed several approaches to measure the distance between the judgess' decision-making processes. Initially, we estimated the distance directly from the dataset. Next, we trained a decision tree model for each judge and measured the distance between the models using Jaccard similarity. Finally, we trained both decision tree and linear regression models for each judge, applied SMOTE

for data augmentation, and calculated the pairwise Mean Squared Error (MSE) between the predictions of these models on the augmented dataset.

5.3 Explain Models and the Differences between Models using LLM

We further analyze the decision-making processes by leveraging Large Language Models (LLMs):

- **Model Explanation**: Train a decision tree model for each judge and use an LLM to explain the decision-making process of each model, focusing on individual judges.
- **Comparative Analysis**: Provide the LLM with pairs of models (representing two judges) and request an identification and explanation of the main differences in their decision-making processes.
- **Textual Descriptions**: Describe the models textually, allowing the LLM to articulate the decision-making patterns and variations between judges.

This approach provides a qualitative understanding of the similarities and differences in clinical judgments among the judges, enhancing the interpretability of the machine learning models.

6 Research Outline

The research conducted so far has provided important insights into the decision-making processes of clinical judges.

6.1 Experimental Overview and Key Findings

In this subsection, we provide a summary of the experiments conducted to analyze the decision-making processes of clinical judges. These experiments include model training, cross-judge predictions, and an analysis of feature importance, all aimed at understanding the variability and consistency in judgments. Below, we outline the key experiments and their findings, providing insights into the effectiveness of different machine learning models and the influence of patient characteristics on decision outcomes.

6.1.1 Single Judge Prediction:

We trained models (Random Forest, Logistic Regression, Linear Regression, XGBoost) to predict a single judge's decisions. This experiment was aimed at understanding how well models could replicate individual judgment patterns. The results, presented in Table 2, show that non-linear models such as Random Forest and XGBoost performed better than linear models, reflecting the complexity of clinical decision-making.

6.1.2 Cross-Judge Prediction:

We trained models on one judge's decisions to predict another's. This cross-prediction analysis allowed us to study the correlation between judges' decision patterns. We observed variability in prediction accuracy, indicating significant differences in judgment approaches. The results of the correlations are presented in Heatmap 1. Judge 2, Judge 5, and Judge 16 appear to have notably high correlations with many others. Judge 16 has consistently high correlation values with almost every other judge, indicating they align well with others.

6.1.3 Aggregated Decision Models:

We trained models on the average and median decisions of the judges to predict test set outcomes. This experiment sought to capture the consensus judgment across multiple judges. Boxplot 2 illustrates the variability in prediction performance, with the median decision providing more consistent results. R-squared score shows how well the model explains the variability in the judge's decisions (0-1 scale). Higher means a better fit. For example value of 0.507272 means that approximately 50.7% of the variance in the judge's decision is explained by that model. Generally, Random Forest has the highest scores, indicating it best captures the judges' decision patterns compared to the other models.

6.1.4 Feature Importance:

We analyzed the influence of patient characteristics on decision outcomes, using SHAP (SHapley Additive exPlanations) to quantify feature importance. SHAP allowed us to identify which features contributed the most to the model's predictions, providing a clear understanding of how each characteristic influenced the decisions. The SHAP values calculated across the test dataset quantify how much each feature influences the model's predictions on average. The higher the SHAP value for a feature, the more important it is in the decision-making process. Features like schizophrenia showed the highest influence on decisions across judges, as demonstrated in Figures 3 and 4. This analysis, enhanced by SHAP, provided insights into which traits judges prioritize in their assessments.

Table 2: A model trained on a single judge, predicting the judge's own decisions on the test set. A warmup to comparing between judge's models.

Model	**Train MSE**	**Test MSE**	**Accuracy**	**Precision**	**Recall**	**F1**
Random Forest	0.117733	0.699422	0.300578	0.284878	0.300578	0.283050
Logistic Regression	0.569767	0.682081	0.317919	0.315165	0.317919	0.303902
Linear Regression	1.212758	1.153373	0.724793	NaN	NaN	NaN
XGBoost	0.187500	0.687861	0.312139	0.306110	0.312139	0.298034

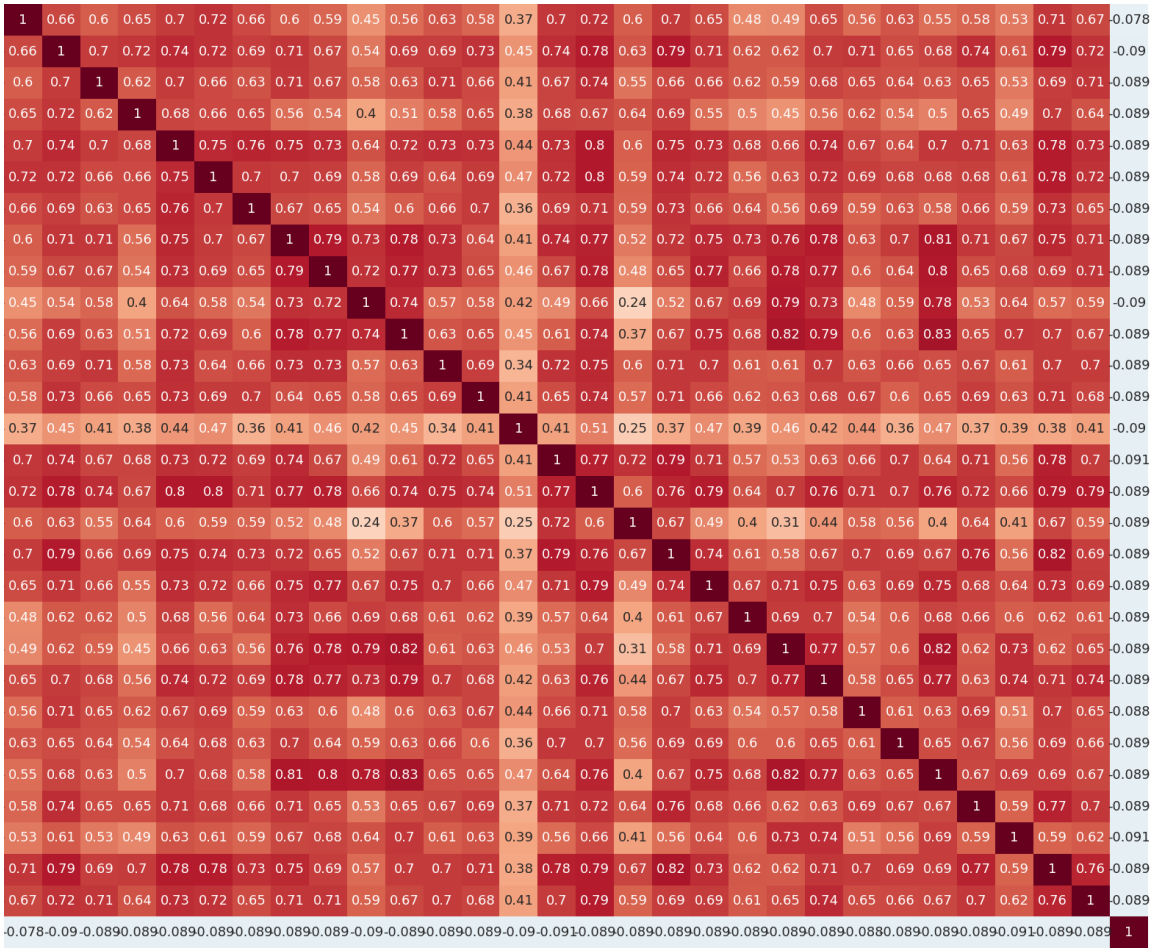

Figure 1: Correlation between the judges presented as a heatmap.

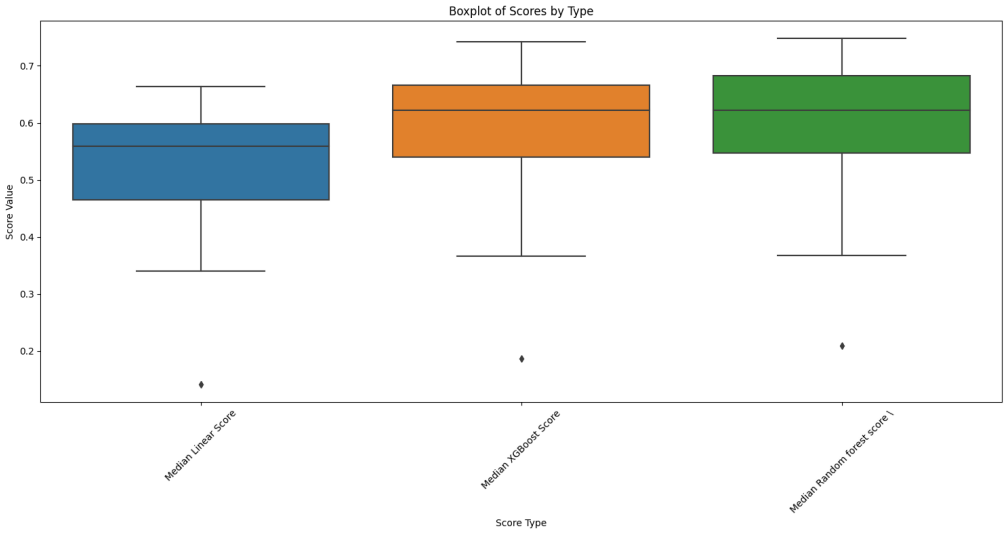

Figure 2: Boxplot representation of the distribution of scores for different scoring methods across multiple judges. The boxplots capture the median, upper, and lower quartiles of the scores, with the whiskers extending to 1.5 times the interquartile range. Each box represents the score distribution for a specific method, indicating the variability and central tendency of the scores provided by the judges.

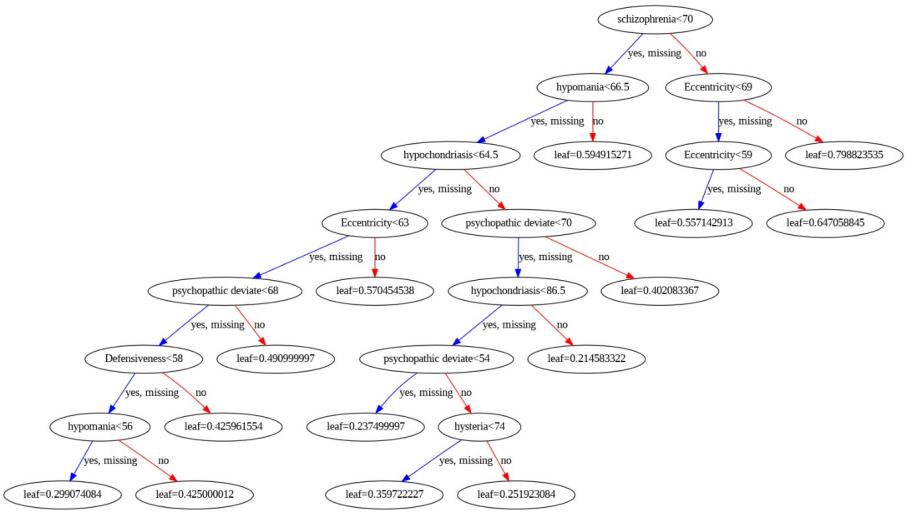

Figure 3: The power of each feature. It can be seen that schizophrenia is the feature with the highest power. Power refers to how much it influences the outcome.

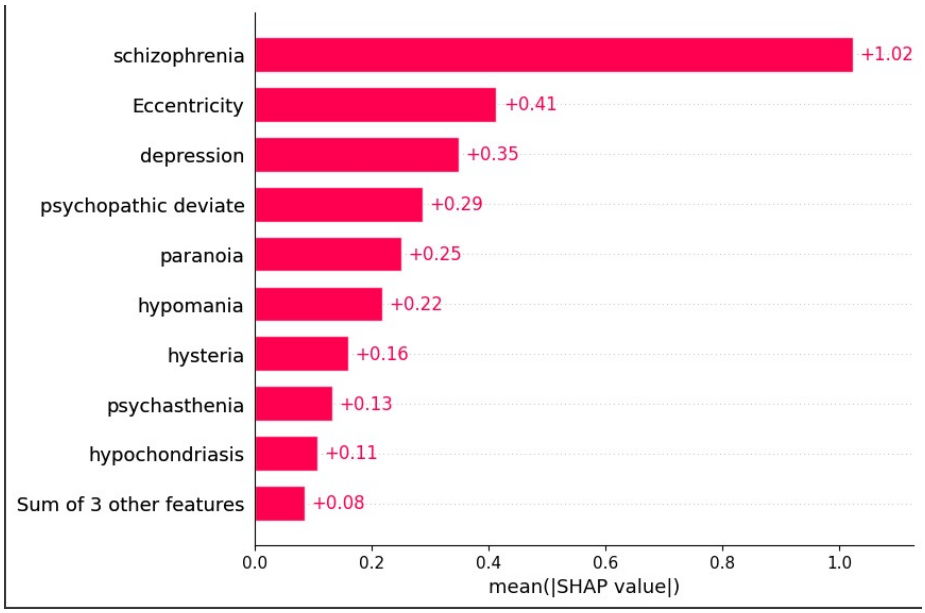

Figure 4: These are the results of the features. Notably, the feature associated with schizophrenia stands out as having the highest mean value

6.2 Estimate the Distance between Judges

Measuring how judges make different decisions is crucial for identifying inconsistencies and improving decision quality. By analyzing the distances between judges' decisions, we can uncover biases, assess training program effectiveness, and enhance teamwork. This analysis reveals patterns in judgment, identifying areas needing improvement.

6.2.1 Naive Distance Estimation

Initially, we estimated the distance directly from the dataset. The results are visualized in a heatmap (Figure 5), where we analyzed the raw differences between judgments.

6.2.2 Decision Tree Model Distance

In our analysis, we measured the distance between the decision-making processes of different judges using the Jaccard similarity index. Specifically, we trained a decision tree model for each judge and then compared the sets of features used by each tree. The Jaccard similarity index, calculated as the ratio of the intersection over the union of the feature sets, provided a measure of similarity between the trees. The tree distance was then defined as $1 - \texttt{jaccard\_similarity}$, with lower distances indicating more similar decision-making processes between

judges. Figure 6 visualizes the dissimilarity matrix, showing how judges vary in their use of patient features.

6.2.3 Distance Estimation on Models with SMOTE

To address potential class imbalances, we applied SMOTE augmentation to the dataset and recalculated the distances. This allowed us to improve model robustness and better reflect judgment similarities and differences.

This approach was implemented using decision tree and linear regression models. The resulting distances are shown in heatmaps for decision trees (Figure 7) and linear regression models (Figure 8).

Data Augmentation and Model Training

For each judge $j \in \{1, 2, \ldots, 29\}$, we trained a decision tree classifier f_j on their respective datasets $\mathcal{D}_j = \{(\mathbf{x}_i, y_{ij})\}_{i=1}^{N}$, where $\mathbf{x}_i \in \mathbb{R}^d$ represents the feature vector of the i-th patient, and $y_{ij} \in \{1, 2, \ldots, 11\}$ denotes the decision made by judge j for patient i.

To ensure robust evaluation, we employed a data augmentation method called Minority Over-sampling Technique Chawla et al. [2002] (SMOTE) to generate augmented datasets $\mathcal{D}_j^{\text{aug}} = \{(\mathbf{x}_{jk}^{\text{aug}}, y_{jk}^{\text{aug}})\}_{k=1}^{M}$, with $M = 5500$. Separate augmented datasets were created for each judge to preserve the integrity of the decision-making patterns.

Distance Calculation

The distance between the decision processes of two judges j and j' was quantified by the Mean Squared Error (MSE) between their predictions on the augmented datasets:

$$\text{MSE}(j, j') = \frac{1}{M} \sum_{k=1}^{2M} (\hat{y}_{kj}^{\text{aug}} - \hat{y}_{kj'}^{\text{aug}})^2$$

where $\hat{y}_{kj}^{\text{aug}}$ and $\hat{y}_{kj'}^{\text{aug}}$ are the predicted decisions for the k-th patient in the augmented dataset by judges j and j', respectively.

A symmetric 29×29 distance matrix $\mathbf{D}$ was constructed, with each element $D_{jj'}$ representing the MSE between the predictions of judges j and j':

$$\mathbf{D} = [D_{jj'}] = [\text{MSE}(j, j')]$$

Results and Discussion

The pairwise MSE heatmaps (Figures 7 and 8) reveal the extent of similarity or dissimilarity in the decision processes of different judges. Lower MSE values indicate higher similarity in decision-making, while higher values suggest greater dissimilarity.

These visualizations provide valuable insights into the consistency and variability of medical decisions across different practitioners. For instance, judges

with similar decision-making patterns cluster together, highlighting potential opportunities for peer learning and standardization of best practices.

A notable observation is the consistently high MSE values associated with **Judge 14**. In the linear regression model (Figure 8), **Judge 14** shows the highest MSE of 4.22 when compared to Judge 17, indicating a significant dissimilarity in decision-making. Similarly, in the decision tree model (Figure 7), **Judge 14** again shows a relatively high MSE of 3.31 when compared to Judge 17. This suggests that Judge 14's decision-making process is markedly different from that of their peers across both models. Such findings highlight the need for further investigation into the underlying reasons for this variability, which may point to differences in judgment criteria, experience, or other factors influencing medical decision-making.

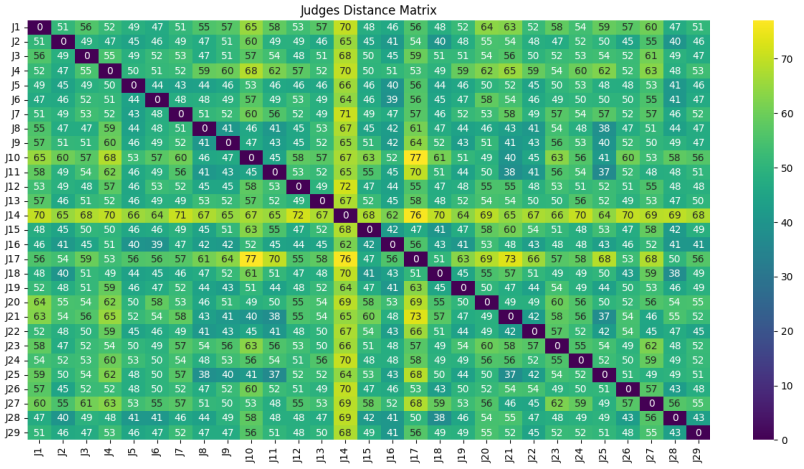

Figure 5: Euclidean distance of the dataset between the judges

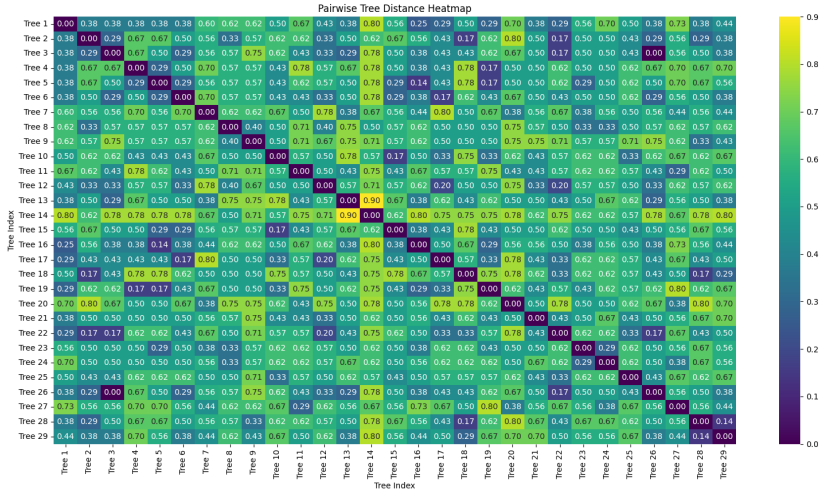

Figure 6: Heatmap of distances between the judges

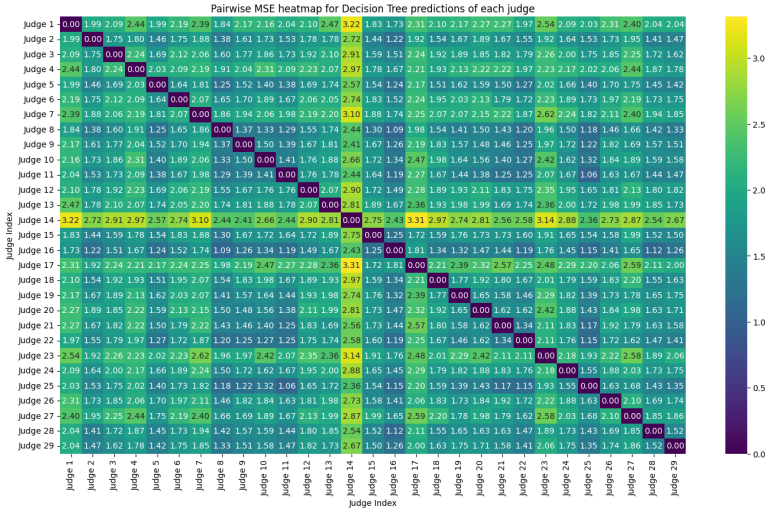

Figure 7: Pairwise MSE heatmap for predictions of each decision tree model of a judge.

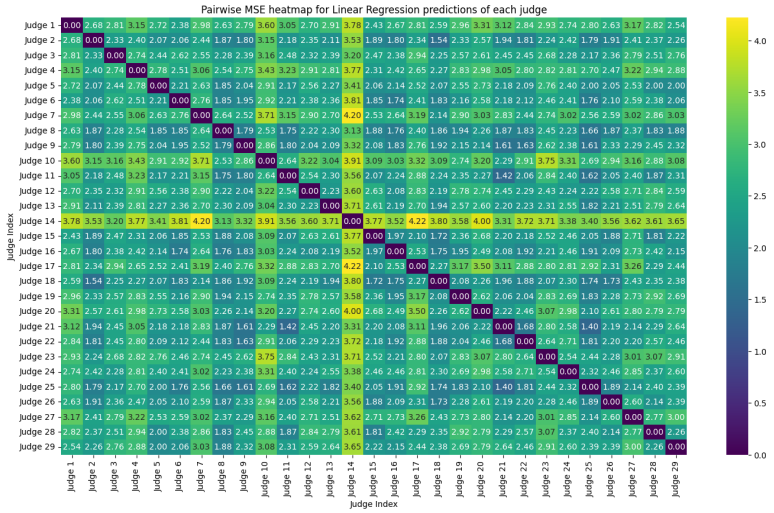

Figure 8: Pairwise MSE heatmap for predictions of each linear regression model of a judge.

6.3 Explainable AI

- **Comparison Using LLMs**: Large Language Models (LLMs) were used to compare the decision-making models of different judges. This comparison helped us identify the underlying patterns and differences between judgments, making the machine learning models more interpretable.
- **Model Evaluation and Interpretation**: By leveraging LLMs, we enhanced the interpretability of decision models, focusing on transparency and clarity in clinical judgments. This made the decision-making processes more understandable to non-experts.
- **Application to Clinical Practice**: The findings from our analysis can be applied to real-world clinical settings, improving the consistency and reliability of medical judgments through the integration of machine learning tools.

Example of Comparative Analysis

6.3.1 Decision Tree Comparisons

To illustrate the differences in decision-making processes, we analyzed the decision trees of two judges. Each model aims to classify a patient's condition (Psychotic/Neurotic) using the features from the MMPI dataset.

Technical Explanation

First Tree:

- **Root Node**: Splits on Defensiveness (defensiveness ¡= 60.5).
- **Key Features**: Hysteria, Paranoia, Depression and Hypomania.
- **Structure**: The tree has a deeper and more varied path, indicating complex decision boundaries.

Second Tree:

- **Root Node**: Also splits on Defensiveness (defensiveness ¡= 60.5).
- **Key Features**: Schizophrenia, Psychopathic Deviate, Hysteria and Hypomania.
- **Structure**: The tree appears more balanced with consistent use of specific features like Schizophrenia and Psychopathic Deviate.

Non-Technical Description
First Judge:

- **Main Focus**: This judge starts by looking at how defensive the patient is. Then, they pay close attention to traits like hysteria, paranoia, and depression. They go through a more complex evaluation, considering a wide range of symptoms to make their decision.
- **Approach**: This judge's approach is thorough, involving many different angles and more in-depth analysis of each symptom.

Second Judge:

- **Main Focus**: Similarly, this judge also begins by evaluating defensiveness but quickly moves to consider schizophrenia and psychopathic traits. They have a more structured approach, focusing on specific symptoms like hysteria and hypomania.
- **Approach**: This judge's method is more straightforward and balanced. They rely on a consistent set of symptoms to form their diagnosis.

Differences Between the Two Judges:

- **First Judge**: Uses a broader variety of symptoms and follows a more intricate path in their diagnostic process, reflecting a detailed and perhaps more cautious approach.
- **Second Judge**: Uses a more consistent and straightforward set of symptoms to make their diagnosis, indicating a balanced and possibly more efficient approach.

Conclusion: Who is the More Experienced Judge?
Based on the differences in the decision trees, the more experienced judge is likely represented by the first tree. Here's why:

- **Complexity and Depth**: The first tree shows a more intricate and varied path in the decision-making process, suggesting that the judge is considering a wider range of symptoms and is taking a more nuanced approach to the diagnosis. An experienced judge is likely to recognize and evaluate subtle patterns and interrelations among different features (symptoms).

- **Thorough Evaluation**: The first tree's thoroughness and attention to multiple features (Defensiveness, Hysteria, Paranoia, Depression, Hypomania) suggest a deep understanding of the condition. An experienced judge is likely to delve deeper into various aspects of the patient's condition, reflecting their comprehensive knowledge and careful consideration.

In contrast, the second tree, while efficient and balanced, indicates a more straightforward approach, which could be indicative of a judge who relies on well-established patterns and might prioritize efficiency over depth.

Thus, the judge represented by the first tree likely has more experience, as they demonstrate a deeper and more nuanced approach to diagnosing the condition.

6.3.2 Linear Regression Comparisons

We also compared the linear regression models of the judges to understand the differences in their decision-making processes from another perspective.

Technical Explanation

Judge 1's Model:

- **Coefficients**:
  - Lie Eccentricity: 0.0089
  - Defensiveness: 0.0239
  - Hypochondriasis: -0.0221
  - Depression: -0.0318
  - Hysteria: 0.0060
  - Psychopathic Deviate: -0.0272
  - Paranoia: 0.0206
  - Psychasthenia: 0.0322
  - Schizophrenia: -0.0413
  - Hypomania: 0.0751

Judge 2's Model:

- **Coefficients**:
  - Lie Eccentricity: 0.0003

- Defensiveness: 0.0512
- Hypochondriasis: 0.0061
- Depression: -0.0068
- Hysteria: -0.0176
- Psychopathic Deviate: -0.0290
- Paranoia: 0.0326
- Psychasthenia: 0.0207
- Schizophrenia: -0.0290
- Hypomania: 0.0863

Non-Technical Description
Judge 1:

- **Main Focus**: This judge considers a variety of symptoms with different weights, indicating a nuanced approach to diagnosis.
- **Approach**: The judge uses a detailed evaluation of multiple features to form their decision.

Judge 2:

- **Main Focus**: This judge also evaluates defensiveness but gives significant importance to schizophrenia and psychopathic deviate traits.
- **Approach**: The judge relies on a consistent set of symptoms with distinct weightings to make their diagnosis.

Differences Between the Two Judges:

- **Judge 1**: Uses a balanced approach with varied symptom weights, reflecting a detailed diagnostic process.
- **Judge 2**: Focuses on specific symptoms with significant weights, indicating a more focused and consistent diagnostic approach.

6.3.3 Conclusion: Integration of Findings

By comparing both decision tree and linear regression models, we gain a comprehensive view of the differences in decision-making processes among judges. The examples provided illustrate how different approaches and symptom evaluations can lead to varied diagnostic processes. For practical implementation, a few representative comparisons are sufficient to highlight these differences effectively.

6.3.4 Scaling the Analysis

Given the dataset of 29 judges, performing a full 29x29 comparative analysis might be excessive. Instead, a few representative examples can effectively illustrate the differences in decision-making processes. For this analysis, the process was done manually through chat interactions with the LLM. For larger datasets, using the OpenAI API to automate the generation of explanations and comparative analyses could streamline the process.

7 Conclusion

In this study, we explored the variability in clinical decision-making among judges using machine learning models. Our experiments revealed several key findings:

7.1 Key Findings

We found that non-linear models such as Random Forest and XGBoost better captured the complexity of individual judgments compared to linear models. Through cross-judge predictions, we identified significant differences in judgment approaches, with variability highlighted in prediction accuracy across judges.

By employing Jaccard similarity and Mean Squared Error (MSE) on augmented datasets, we quantified the distances between judges' decision-making processes. The distance analysis revealed clusters of judges with similar decision patterns, suggesting opportunities for peer learning and improving decision consistency. A notable outlier, Judge 14, exhibited significant dissimilarity in both decision tree and linear regression models, warranting further investigation.

7.2 Variability in Clinical Judgments

The observed variability in judgments can be attributed to several factors. First, judges may interpret patient features like schizophrenia and paranoia differently, leading to varied decisions. Second, differences in experience levels result in distinct approaches—experienced judges may use more nuanced, in-depth evaluations, while less experienced judges rely on streamlined decision patterns. Finally, biases and training differences contribute to how judges prioritize symptoms, further exacerbating variability.

7.3 Role of Explainability

Additionally, we used Large Language Models (LLMs) to provide transparent explanations of decision-making patterns. These explanations allowed us to compare the complexity and depth of decisions made by different judges, offering

insights into how expertise and judgment style influence variability. The integration of machine learning and LLM-based explainability presents a promising approach to enhancing the interpretability of clinical judgments.

7.4 Implications for Practice

Our findings demonstrate that understanding variability in clinical decision-making can provide valuable insights into improving consistency, training, and standardization in clinical practice. Future research could extend this work by automating the comparison and explanation processes for larger datasets, enabling a more scalable approach to analyzing expert judgment variability.

7.5 Future Work

Future research could focus on several key areas:

- **Automation of Analysis:** Developing automated systems using OpenAI APIs to streamline the generation of explanations and comparative analyses for larger datasets. This would enhance efficiency and scalability in analyzing clinical judgments.
- **Broader Datasets:** Applying this methodology to larger and more diverse datasets to assess the generalizability of findings across different clinical settings and populations.
- **Longitudinal Studies:** Conducting longitudinal studies to track how judges' decision-making processes evolve over time, especially after targeted training interventions.
- **Integration of Other Models:** Exploring the integration of additional machine learning models, such as ensemble methods or neural networks, to capture more complex decision-making patterns.
- **Feedback Mechanisms:** Implementing feedback mechanisms where judges can review and reflect on model predictions and explanations, fostering continuous learning and improvement in clinical practice.

These directions will not only enhance our understanding of clinical decision-making but also contribute to more effective training and support for clinicians.

8 References

Lorenz Adlung, Yotam Cohen, Uria Mor, and Eran Elinav. Machine learning in clinical decision making. Frontiers in Medicine, 8, 2022. doi: 10.3389/fmed.2021.765693.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16:321–357, 2002.

Yoav Ganzach. Nonlinear models of clinical judgment: Communal nonlinearity and nonlinear accuracy. Psychological Science, 12(5):403–407, 2001. doi: 10.1111/1467-9280.00374. URL https://doi.org/10.1111/1467-9280.00374. PMID: 11554674.

Paul E. (Paul Everett) Meehl. Clinical versus statistical prediction : a theoretical analysis and a review of the evidence. University of Minnesota Press, Minneapolis. URL https://doi.org/10.1037/11281-000.

Tim Miller et al. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267:1–38, 2017. URL https://doi.org/10.1016/j.artint.2018.07.007.

Abby L. Mulay, Mark Waugh, Michael Finn, Jenna Gilmore, Marisa Whitley, Nicole Cain, Robert Gordon, Alvin Jones, David Nichols, Gina Rossi, and David Streiner. Mmpi-2 personality disorder spectra scales. pages 3–41, 01 2019. URL https://www.researchgate.net/publication/330263734_MMPI-2_Personality_Disorder_Spectra_Scales.