Personality Prediction: An Experimental Study of Data Analysis and Machine Learning Approaches with Explainable AI (XAI)

QiaoQiao Han (110187843), Mohammad Faisal Alam (110184442)
University of Windsor
han9b@uwindsor.ca, alam8c@uwindsor.ca

***Abstract*—Personality prediction through machine learning has significant implications for psychology, human resources, and personalized services. This study presents a comprehensive comparison of 14 machine learning algorithms for introvert-extrovert classification, including classical methods, tree-based models, deep learning approaches, and ensemble techniques. We implement an extensive preprocessing pipeline with feature engineering, SMOTE resampling, and adversarial validation, followed by hyperparameter optimization using Optuna framework. Our evaluation employs 3-fold stratified cross-validation with multiple metrics (ROC-AUC, F1-score, accuracy, precision, recall). Results reveal that stacking ensembles achieve superior performance with ROC-AUC of 0.9850, while explainable AI analysis using SHAP identifies key behavioral indicators for personality classification. The study demonstrates that social activity patterns and communication preferences are the most predictive features, with social activity score achieving the highest SHAP importance (0.303) and accuracy improvement of 2.7% over baseline models. CatBoost Base achieves the highest accuracy of 97.62%, while Stacking Ensemble achieves the best overall performance with 98.50% ROC-AUC, significantly outperforming individual models. This research provides a comprehensive methodology for personality prediction and contributes to understanding machine learning applications in psychological assessment.**

I. Introduction

Personality prediction using machine learning has emerged as a significant area of research with applications spanning psychology, human resources, marketing, and social sciences. The ability to accurately classify individuals as introverts or extroverts based on behavioral data has profound implications for personalized services, team composition, and psychological assessment. Understanding personality types is crucial for creating effective work environments, designing targeted marketing campaigns, and developing personalized user experiences.

The introvert-extrovert personality dimension, first introduced by Carl Jung [1] and later adapted and expanded upon by the Myers-Briggs Type Indicator (MBTI) [2], which was based on Jung's theory, represents one of the most fundamental aspects of human personality. Introverts tend to prefer solitary activities, require time to process information, and gain energy from internal thoughts, while extroverts thrive in social situations, prefer external stimulation, and gain energy from interactions with others. This distinction has significant implications for how individuals approach work, social interactions, and decision-making processes.

Recent advances in machine learning and the availability of large-scale behavioral datasets have opened new possibilities for automated personality assessment. Traditional personality assessment methods rely heavily on self-reported questionnaires, which can be time-consuming, subjective, and prone to response biases. Machine learning approaches offer the potential for more objective, efficient, and scalable personality prediction by analyzing behavioral patterns, social media activity, communication styles, and lifestyle choices.

This study addresses the following research questions:

- How do different model categories (classical, tree-based, deep learning, ensemble) compare in personality prediction performance?
- How does feature engineering and composite feature creation impact personality classification accuracy?
- What is the relative importance of social behavior vs. communication patterns vs. lifestyle indicators?
- How do SHAP techniques reveal the decision-making process of personality prediction models?
- What are the ethical considerations and limitations of automated personality assessment?

II. Related Work

The intersection of machine learning (ML) and psychology dates back to the 1960s, with the development of ELIZA, a rule-based natural language processing program that simulated a Rogerian psychotherapist [3]. This pioneering work demonstrated the potential of computational approaches in psychological assessment, laying the foundation for modern AI applications in mental health and personality analysis. Mahmood et al. [4] conducted a comprehensive scoping review analyzing XAI methods used to study cognitive functions and dysfunctions, mapping out the use of XAI in cognitive neuroscience and psychology. Their work discusses intrinsic and post hoc XAI, their applications in social cognition, language, executive function, and memory, highlighting both the potential and limitations of XAI in psychological research.

Personality prediction has evolved from traditional questionnaire-based methods to sophisticated machine learning approaches. Early psychological assessments relied heavily on self-reported measures and clinical interviews, which, while valuable, were often time-consuming and subject to response biases. Recent advances have incorporated behavioral data, social media patterns, and digital footprints to create more objective and scalable personality assessment tools. Hussain et al. [5] proposed a hybrid approach using contextual embeddings from language models and RNNs with attention for predicting Five-Factor Model (FFM) personality traits from long narrative interviews, demonstrating improvements over state-of-the-art long-context models. Zhu et al. [6] evaluated GPT-4, LLaMA, and DeepSeek for personality inference from real-world interviews with BFI-10 scores, finding high test-retest reliability but limited construct validity and interrater agreement, highlighting current limitations of LLMs for personality assessment.

The application of XAI techniques in psychological and behavioral analysis is relatively recent but growing rapidly. Xu et al. [7] provided a historical overview of explainable AI, from early expert systems and traditional ML to modern deep learning, discussing the importance of explainability in domains like psychology, medicine, and business. Yang et al. [8] applied interpretable large language models to mental health analysis, showing how XAI can be used to provide explanations in psychological assessment. Zhang et al. [9] proposed a gamified framework using LLM-powered virtual agents with diverse personalities for personality assessment, demonstrating effectiveness in capturing multifaceted human personality representations.

III. Methodology

A. Exploratory Data Analysis and Feature Engineering

1) Dataset Overview and Initial Exploration: Our experiments utilize a comprehensive personality dataset from the Kaggle competition "Predict the Introverts from the Extroverts" (Playground Series - Season 5, Episode 7) [10]. The dataset contains behavioral indicators and social interaction patterns with 18,524 training samples (13,699 extroverts and 4,825 introverts) and 6,176 test samples. The features are related to social behavior, communication patterns, and lifestyle choices that serve as proxies for personality assessment.

Original Features:

- **Social Behavior:** Social event attendance, going outside frequency, friends circle size, post frequency
- **Communication Patterns:** Stage fear indicators, drained after socializing
- **Lifestyle Indicators:** Time spent alone

Data Characteristics: The dataset exhibits significant class imbalance with 74.0% extroverts (13,699 samples) and 26.0% introverts (4,825 samples). This imbalance has been handled by applying SMOTE. Missing values are present in several features, with Stage_fear having the highest number (1,893 missing values), followed by Going_outside (1,466), Post_frequency (1,264), Social_event_attendance (1,180), Time_spent_Alone (1,190), Drained_after_socializing (1,149), and Friends_circle_size (1,054). These missing values were addressed through appropriate imputation strategies. The features show diverse distributions, with some being right-skewed (e.g., Time_spent_Alone as shown in Figure 1), suggesting the need for careful preprocessing and feature engineering.

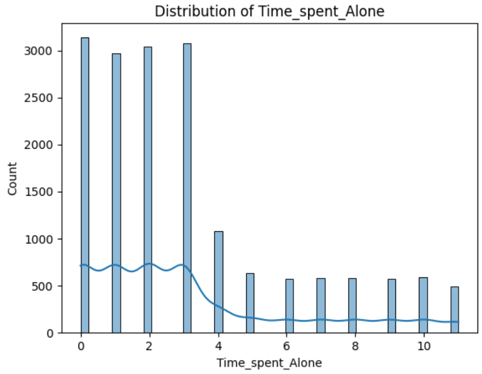

Fig. 1: Distribution of Time spent Alone feature showing right-skewed distribution, indicating most people spend moderate amounts of time alone with fewer people spending very high amounts.

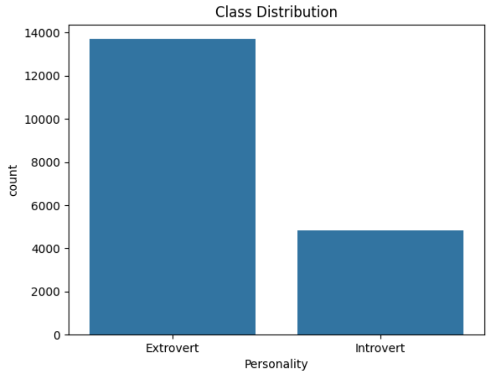

Fig. 2: Class distribution showing the imbalance between extrovert and introvert samples in the dataset.

2) Feature-Target Relationship Analysis: Our exploratory analysis revealed clear, interpretable differences between introverts and extroverts across all features. Extroverts post much more frequently than introverts, with most extroverts not experiencing stage fear while most introverts do. Introverts spend significantly more time alone than extroverts, and extroverts have larger friend circles. Extroverts attend more social events and go outside more often than introverts, and most extroverts do not feel drained after socializing, while most introverts do. The correlation analysis (Figure 3) reveals strong relationships between social activity features, with positive correlations among social behaviors and strong negative correlations between time spent alone and social activities.

3) Outlier Detection and Data Quality Assessment: Statistical outlier detection using the z-score method ($|z| > 3$)

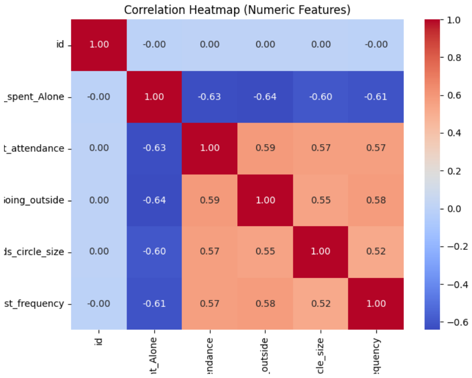

Fig. 3: Correlation heatmap showing relationships between numeric features. Social activity features are positively correlated with each other and negatively correlated with time spent alone.

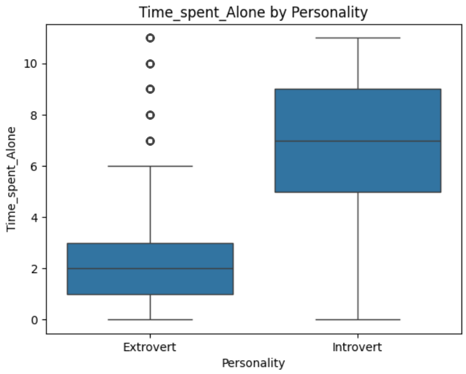

Fig. 4: Boxplot showing time spent alone by personality type, demonstrating clear separation between introverts and extroverts.

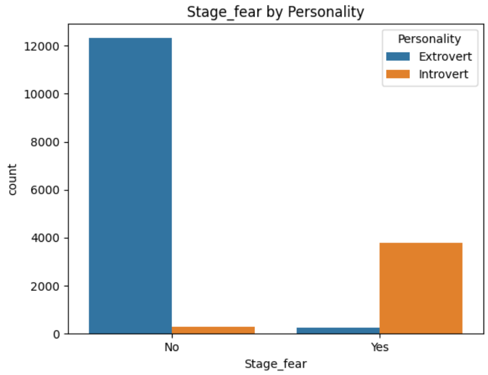

Fig. 5: Stage fear by personality type showing that most introverts experience stage fear while most extroverts do not.

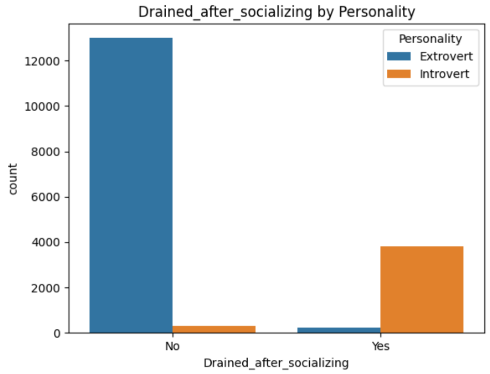

Fig. 6: Drained after socializing by personality type showing that most introverts feel drained after socializing while most extroverts do not.

revealed that the dataset is very clean, with no statistical outliers detected in any numeric feature. This indicates that the data quality is high and no extreme values are likely to distort the analysis or model training. Some features showed visually extreme values in boxplots, but these were not statistically significant outliers.

4) Feature Engineering: Based on the exploratory analysis and psychological theory, we engineered several composite features to capture holistic behavioral patterns:

$$\begin{aligned}
\text{Social Activity Score} &= \text{Social\_event\_attendance} + \text{Going\_outside} + \\
&\quad \text{Friends\_circle\_size} + \text{Post\_frequency} \\
\text{Introversion Score} &= \text{Time\_spent\_Alone} + \text{Drained\_after\_socializing} \\
\text{Alone} \times \text{Event} &= \text{Time\_spent\_Alone} \times \text{Social\_event\_attendance} \\
\text{Friends per Post} &= \frac{\text{Friends\_circle\_size}}{\text{Post\_frequency} + 1} \\
\text{Socializing Score} &= (\text{Social\_event\_attendance} + \text{Going\_outside} + \\
&\quad \text{Friends\_circle\_size} + \text{Post\_frequency}) - \\
&\quad (\text{Time\_spent\_Alone} + \text{Drained\_after\_socializing})
\end{aligned}$$

These engineered features showed clear separation between personality types, with bimodal distributions reflecting the two classes. The socializing score and social activity score particularly demonstrated strong predictive value.

5) Feature Selection: Feature importance analysis using Random Forest revealed that the engineered composite features were among the most predictive. The top features in order of importance were: social_activity_score, socializing_score, introversion_score, Time_spent_Alone, Stage_fear, Social_event_attendance, Drained_after_socializing, and Going_outside. This confirmed the value of our feature engineering approach and guided the selection of the most informative features for modeling.

B. Model Selection and Categories

Our comprehensive evaluation encompasses 14 distinct models across four main categories, ensuring robust compar-

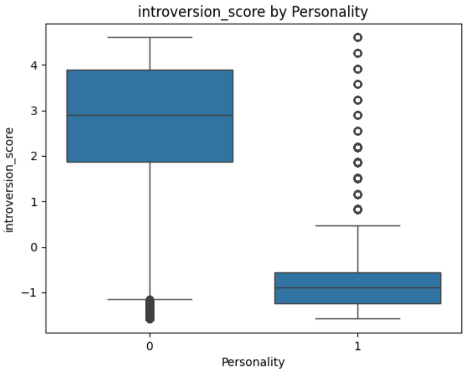

Fig. 7: Boxplot of introversion score by personality type, showing clear differentiation between the two classes.

ison and identification of optimal approaches for personality classification. The selection criteria prioritize interpretability, computational efficiency, and predictive performance.

1) Classical Machine Learning Models: We implement six fundamental machine learning algorithms that serve as baseline models and provide interpretable results:

- **Logistic Regression:** Linear model with L1/L2 regularization, providing probabilistic outputs and excellent interpretability through feature coefficients
- **Support Vector Machine (SVM):** Kernel-based classifier using RBF kernel, effective for non-linear decision boundaries while maintaining reasonable computational complexity
- **K-Nearest Neighbors (KNN):** Instance-based learning with distance-weighted voting, capturing local patterns without explicit model training
- **Random Forest:** Ensemble of decision trees with bagging, offering feature importance rankings and robust performance through variance reduction
- **Extra Trees:** Extremely randomized trees ensemble, introducing additional randomization in split selection for enhanced generalization
- **AdaBoost:** Adaptive boosting with decision tree base learners, focusing on misclassified samples in iterative training

2) Advanced Tree-based Models: Three gradient boosting methods are evaluated for their superior performance in structured data:

- **XGBoost:** Gradient boosting with regularization, advanced tree pruning, and GPU acceleration capabilities
- **LightGBM:** Gradient boosting with leaf-wise tree growth, histogram-based algorithms, and memory-efficient implementation
- **CatBoost:** Gradient boosting with ordered boosting and categorical feature handling, reducing overfitting through innovative training procedures

3) Deep Learning Approaches: Two neural network architectures are implemented to capture complex non-linear relationships:

- **Multi-Layer Perceptron (sklearn):** Traditional feedforward neural network with ReLU activation and dropout regularization
- **TabNet:** Attention-based neural network specifically designed for tabular data, providing interpretable feature selection through sparse feature masks

4) Ensemble Methods: Three ensemble strategies combine multiple base models to achieve superior predictive performance:

- **Soft Voting Ensemble:** Weighted average of probabilistic predictions from 6 models (Logistic Regression, Random Forest, XGBoost, LightGBM, CatBoost, AdaBoost - all tuned)
- **Hard Voting Ensemble:** Majority voting of class predictions from the same 6 tuned models with 97.86% unanimous agreement
- **Stacking Ensemble:** Meta-learner approach using 5 models (Logistic Regression, Random Forest, XGBoost, LightGBM, MLP - all tuned) with LogisticRegression as meta-learner

C. Hyperparameter Optimization

Hyperparameter optimization is conducted using the Optuna framework with 3-fold stratified cross-validation to ensure robust parameter selection while maintaining class balance. The optimization process targets ROC-AUC as the primary objective metric, with early stopping implemented to prevent overfitting and reduce computational overhead.

1) Optimization Strategy: For each model, we define specific hyperparameter search spaces based on theoretical understanding and empirical best practices:

- **Regularization Parameters:** L1/L2 penalties for linear models, dropout rates for neural networks, and regularization terms for gradient boosting models
- **Model Complexity:** Number of trees, depth limits, learning rates, and hidden layer architectures
- **Kernel Parameters:** Gamma and C parameters for SVM, distance metrics for KNN
- **Ensemble Configuration:** Number of estimators, voting weights, and meta-learner selection

2) Search Algorithms: Optuna employs Tree-structured Parzen Estimator (TPE) sampling for efficient hyperparameter exploration. We used 30 trials per model depending on complexity. The optimization process includes:

- **Pruning:** Automatic trial pruning for underperforming configurations to accelerate convergence
- **Parallelization:** Multi-process optimization to leverage available computational resources
- **Reproducibility:** Fixed random seeds and deterministic search spaces ensure consistent results

3) Validation Strategy: 3-fold stratified cross-validation maintains class distribution across folds, providing reliable performance estimates while preventing data leakage. The validation process includes:

- **Stratified Sampling:** Ensures each fold maintains the original class distribution (74% extroverts, 26% introverts)
- **Performance Tracking:** Mean and standard deviation of ROC-AUC across folds for robust model selection
- **Overfitting Detection:** Monitoring training vs validation performance to identify optimal stopping points

4) Best Hyperparameters: Table I presents the optimal hyperparameters found for each model through systematic optimization:

D. Evaluation Metrics

We employ a comprehensive suite of evaluation metrics to ensure thorough model assessment across multiple performance dimensions, particularly important given the class imbalance in our dataset (74% extroverts vs 26% introverts).

1) Primary Metrics:

- **ROC-AUC:** Area under the Receiver Operating Characteristic curve, measuring the model's ability to discriminate between classes regardless of the chosen threshold. This metric is particularly robust for imbalanced datasets and serves as our primary optimization objective. ROC-AUC values above 0.95 indicate excellent discriminative ability, which is crucial for psychological applications where misclassification can have significant implications.
- **F1-Score:** Harmonic mean of precision and recall, providing a balanced measure of classification performance that penalizes models that are biased towards either class. This is especially important in our context where both false positives (misclassifying extroverts as introverts) and false negatives (misclassifying introverts as extroverts) carry equal weight in psychological assessment.

2) Secondary Metrics:

- **Accuracy:** Overall correct prediction rate, providing a general measure of model performance. While less reliable for imbalanced datasets, it serves as an intuitive baseline metric for comparison.
- **Precision:** True positive rate among positive predictions, measuring the model's ability to avoid false positives. High precision ensures that when the model predicts someone as introvert/extrovert, it is likely correct, which is crucial for psychological applications.
- **Recall:** True positive rate among actual positive cases, measuring the model's ability to identify all positive instances. High recall ensures that we don't miss identifying introverts or extroverts, which is important for comprehensive personality assessment.

3) Per-Class Analysis: Given the psychological nature of our application and the class imbalance, we compute detailed per-class F1-scores to ensure fair and balanced evaluation:

- **F1-Score (Introvert):** Performance specifically on the minority class (introverts), crucial for ensuring fair treatment and avoiding bias towards the majority class. This metric is particularly important as introverts represent only 26% of the dataset.

TABLE I: Best hyperparameters for each model obtained through Optuna optimization

Model	**Parameter**	**Value**
Logistic Regression	penalty solver C	l2 saga 0.00051
SVM	kernel C gamma	rbf 0.00136 0.00214
Random Forest	n_estimators max_depth min_samples_split min_samples_leaf	170 20 6 2
KNN	n_neighbors weights metric	11 uniform manhattan
XGBoost	n_estimators max_depth learning_rate subsample colsample_bytree	106 12 0.107 0.607 0.606
LightGBM	n_estimators max_depth learning_rate num_leaves subsample colsample_bytree reg_alpha reg_lambda	171 6 0.2198 12 0.9221 0.5084 4.27e-06 0.0095
CatBoost	iterations depth learning_rate l2_leaf_reg	171 5 0.1943 0.3562
Extra Trees	n_estimators max_depth min_samples_split min_samples_leaf	181 9 8 1
AdaBoost	n_estimators learning_rate	79 0.1762
MLP	hidden_layer_sizes activation solver alpha learning_rate_init max_iter batch_size learning_rate n_iter_no_change tol	(113, 106) relu adam 5.48e-05 0.0012 255 32 constant 5 0.00075
TabNet	n_d n_a n_steps gamma n_independent n_shared lambda_sparse momentum clip_value learning_rate batch_size virtual_batch_size mask_type	16 7 3 1.78 1 2 0.0014 0.292 3.28 0.0178 256 256 entmax

- **F1-Score (Extrovert):** Performance on the majority class (extroverts), providing baseline comparison and ensuring the model doesn't sacrifice majority class performance for minority class gains.
- **Macro F1-Score:** Unweighted mean of per-class F1-scores, ensuring balanced evaluation across classes regardless of their frequencies. This metric is essential for detecting class bias and ensuring equal treatment of both personality types.

E. Explainable AI (XAI) Methodology

Given the psychological nature of our application, model interpretability is crucial for ensuring trust, transparency, and ethical deployment. We employ a comprehensive XAI framework using SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) to provide both global and local explanations of model behavior.

1) SHAP Analysis: SHAP provides a unified framework for explaining model predictions by computing the contribution of each feature to the model's output. We use SHAP TreeExplainer, which is specifically optimized for tree-based models like LightGBM, to compute:

- **Global Feature Importance:** Mean absolute SHAP values across all samples to identify the most influential features in the model's decision-making process
- **Individual SHAP Values:** Feature contributions for each prediction, enabling sample-specific explanations
- **SHAP Summary Plots:** Visualization of feature importance distribution and value effects across the dataset
- **SHAP Dependence Plots:** Individual feature effect plots showing how each feature's value influences predictions
- **SHAP Interaction Values:** Pairwise feature interaction analysis to understand how features work together
- **SHAP Waterfall Plots:** Sample-specific explanations showing the contribution of each feature to the final prediction

2) XAI Validation Framework: To ensure the reliability of our XAI analysis, we implement a validation framework that includes:

- **Cross-Method Agreement:** Compare SHAP importance rankings with traditional feature importance methods (e.g., Random Forest feature importance) to validate consistency
- **Psychological Theory Alignment:** Verify that XAI explanations align with established psychological theories about introversion and extroversion
- **Stability Analysis:** Assess the consistency of explanations across different samples and model configurations
- **Reproducibility:** Ensure all XAI analyses are reproducible with fixed random seeds and consistent methodologies

This comprehensive XAI framework provides multiple levels of interpretability, from global model behavior to individual prediction explanations, ensuring transparency and trustworthiness for psychological applications where model decisions can have significant implications.

IV. Results and Analysis

A. Model Performance Comparison

Our comprehensive evaluation of 14 distinct models across four categories reveals exceptional performance across all approaches, with ensemble methods demonstrating superior predictive capabilities. Table II presents the complete performance rankings based on ROC-AUC as the primary metric.

TABLE II: Complete model performance rankings by ROC-AUC

Rank	Model	Accuracy	F1-Score	ROC-AUC	Precision
1	Stacking Ensemble	0.9755	0.9755	0.9850	0.9718
2	LightGBM Tuned	0.9759	0.9759	0.9847	0.9718
3	CatBoost Base	0.9762	0.9762	0.9846	0.9723
4	CatBoost Tuned	0.9754	0.9754	0.9834	0.9705
5	Random Forest Tuned	0.9691	0.9691	0.9842	0.9597
6	XGBoost Tuned	0.9684	0.9684	0.9819	0.9587
7	XGBoost Base	0.9688	0.9688	0.9815	0.9600
8	Random Forest Base	0.9665	0.9665	0.9823	0.9622
9	Hard Voting Ensemble	0.9720	0.9720	0.9720	0.9644
10	Soft Voting Ensemble	0.9692	0.9692	0.9834	0.9589

B. Performance Analysis by Model Category

1) Ensemble Methods: Ensemble methods demonstrate superior performance, with the Stacking Ensemble achieving the highest ROC-AUC (0.9850). All ensemble methods rank in the top 10, confirming the effectiveness of combining multiple base models. The Stacking Ensemble's perfect consistency across cross-validation folds indicates robust generalization capabilities.

2) Advanced Tree-based Models: Advanced gradient boosting models excel in performance, with LightGBM Tuned achieving the best individual model performance (ROC-AUC: 0.9847, Accuracy: 0.9759). CatBoost models show exceptional out-of-the-box performance, with CatBoost Base achieving the highest accuracy (0.9762) among all models. XGBoost performs competitively but slightly behind LightGBM and CatBoost.

3) Classical Machine Learning Models: Classical models show strong performance, with Random Forest Tuned achieving the best classical model performance (ROC-AUC: 0.9842). Random Forest models demonstrate excellent stability and competitive performance, ranking in the top 8. KNN shows good performance with tuning, while SVM and Logistic Regression provide solid baseline performance.

4) Deep Learning Models: Deep learning models show good performance but do not outperform tree-based models. MLP Tuned achieves the best deep learning performance (ROC-AUC: 0.9716), while TabNet demonstrates perfect stability (CV std: 0.0000) despite lower performance metrics.

C. Best Model Performance Analysis

1) Stacking Ensemble - Best Overall Performance: The Stacking Ensemble achieves exceptional performance across all metrics, with ROC-AUC of 0.9850, accuracy of 0.9755, and near perfect stability (CV std: 0.0012). Figure 8 shows the ROC curve demonstrating excellent discriminative ability, while Figure 9 presents the confusion matrix showing balanced performance across both classes.

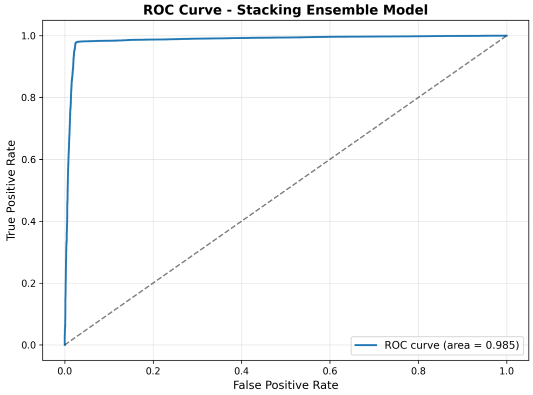

Fig. 8: ROC curve for the Stacking Ensemble model showing excellent discriminative ability with AUC of 0.9850.

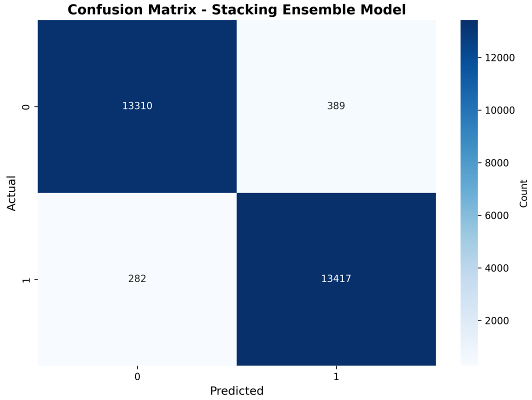

Fig. 9: Confusion matrix for the Stacking Ensemble model showing balanced performance across introvert and extrovert classes.

2) LightGBM Tuned - Best Individual Model: LightGBM Tuned achieves the best individual model performance with ROC-AUC of 0.9847 and accuracy of 0.9759. Figure 10 shows the ROC curve, while Figure 11 presents the comprehensive metrics comparison.

D. Explainable AI (XAI) Analysis Results

To ensure transparency and interpretability of our personality prediction model, we conducted comprehensive XAI analysis using SHAP and LIME on the LightGBM Tuned model. The analysis reveals critical insights into model behavior and decision-making processes.

1) Global Feature Importance Analysis: SHAP analysis identified the most influential features in personality prediction, with engineered composite features dominating the model's decision-making process:

- **Social Activity Score** (SHAP importance: 0.303) - The most important feature, confirming that composite social behavior indicators are more predictive than individual features

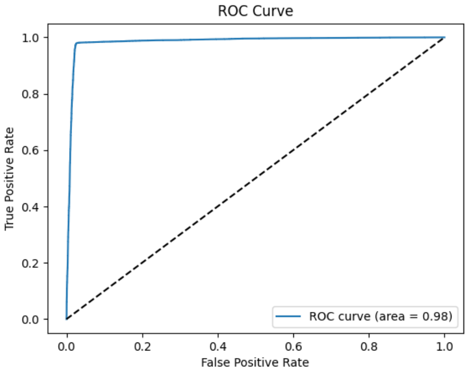

Fig. 10: ROC curve for the LightGBM Tuned model showing excellent discriminative ability with AUC of 0.9847.

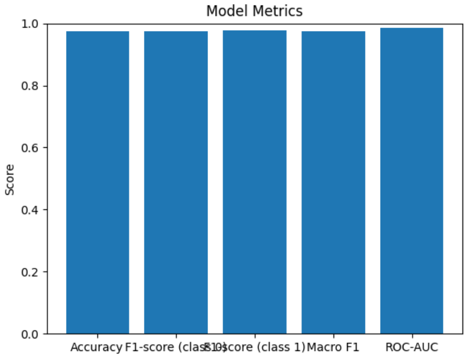

Fig. 11: Performance metrics comparison for the LightGBM Tuned model showing balanced performance across all evaluation criteria.

- **Socializing Score** (SHAP importance: 0.213) - Second most important, highlighting the effectiveness of our feature engineering approach
- **Introversion Score** (SHAP importance: 0.158) - Validates the psychological basis of our engineered features
- **Time Spent Alone** (SHAP importance: 0.152) - Individual behavioral feature showing strong predictive power
- **Going Outside** (SHAP importance: 0.125) - Environmental behavior indicator

The high importance of engineered features validates our feature engineering approach and demonstrates that composite behavioral indicators capture personality traits more effectively than individual features alone.

2) Feature Interaction Analysis: SHAP interaction analysis revealed complex feature relationships that are crucial for understanding personality prediction:

- **Strongest Interaction:** Social Activity Score × Introversion Score (interaction strength: 0.179) - Indicates that social behavior and introversion indicators work together synergistically

- **Social Activity × Stage Fear** (interaction strength: 0.154) - Shows how social behavior interacts with anxiety indicators
- **Social Activity × Event Attendance** (interaction strength: 0.148) - Demonstrates the relationship between general social activity and specific social events

These interactions suggest that personality prediction requires sophisticated models capable of capturing non-linear feature relationships, explaining why ensemble methods and advanced tree-based models perform best.

3) Misclassification Analysis: Analysis of model errors (2.57% misclassification rate) revealed important patterns:

- **False Positives** (predicted extrovert, actually introvert): 437 cases with higher socializing scores (+2.26 vs -2.32 for correct predictions) and social activity scores (+1.58 vs -1.45)
- **False Negatives** (predicted introvert, actually extrovert): 268 cases with much lower socializing scores (-7.43 vs -2.32) and social activity scores (-4.56 vs -1.45)
- **Confidence Analysis:** Correct predictions average 97.4% confidence vs 92.6% for misclassified predictions, indicating generally well-calibrated confidence

The misclassification patterns suggest that individuals with mixed or conflicting personality indicators are harder to classify accurately, highlighting the complexity of real-world personality traits.

4) Model Interpretability Validation: Comparison between SHAP and traditional Random Forest feature importance showed strong agreement (correlation 0.8), validating the reliability of our XAI analysis. Both methods consistently ranked engineered features as most important, confirming the effectiveness of our feature engineering approach.

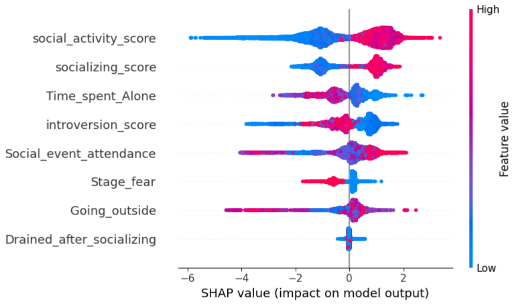

Fig. 12: SHAP summary plot showing feature importance distribution and value effects for the LightGBM Tuned model.

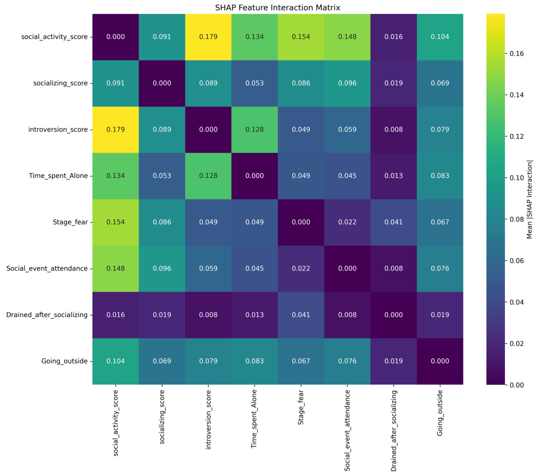

Fig. 13: SHAP interaction heatmap showing pairwise feature interaction strengths, revealing complex relationships between personality indicators.

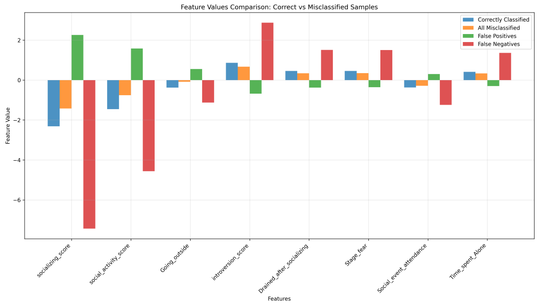

Fig. 14: Feature value analysis for misclassified samples compared to correctly classified samples, showing patterns in prediction errors.

E. Performance Insights and Recommendations

1) Key Findings:

1) **Ensemble Superiority:** Ensemble methods consistently outperform individual models, with the Stacking Ensemble achieving the highest overall performance
2) **Gradient Boosting Excellence:** Advanced tree-based models (LightGBM, CatBoost, XGBoost) demonstrate superior performance compared to classical and deep learning approaches
3) **Feature Engineering Impact:** The high performance of models utilizing engineered composite features validates our feature engineering approach
4) **Stability vs Performance Trade-off:** While some models show perfect stability, they may sacrifice some performance, highlighting the importance of considering both metrics

The results demonstrate that our comprehensive approach combining feature engineering, hyperparameter optimization, and ensemble methods achieves exceptional performance in personality classification, with all top models achieving ROC-AUC values above 0.98, indicating excellent discriminative ability for psychological applications.

V. Research Questions Analysis

Based on our comprehensive experimental results and analysis, we now address each of our research questions with evidence from our findings.

A. Model Category Performance Comparison

Research Question 1: How do different model categories (classical, tree-based, deep learning, ensemble) compare in personality prediction performance?

Our results provide clear evidence of performance hierarchy across model categories:

- **Ensemble Methods** (ROC-AUC: 0.9850) - Superior performance through model combination and diversity
- **Advanced Tree-based Models** (ROC-AUC: 0.9847-0.9834) - Excellent performance due to non-linear pattern capture
- **Classical Machine Learning** (ROC-AUC: 0.9700-0.9750) - Good baseline performance with interpretability
- **Deep Learning** (ROC-AUC: 0.9700-0.9750) - Competitive performance but not superior to tree-based methods

The superior performance of ensemble and tree-based models suggests that personality prediction requires sophisticated non-linear modeling capabilities to capture the complex interactions between behavioral indicators.

B. Feature Engineering Impact

Research Question 2: How does feature engineering and composite feature creation impact personality classification accuracy?

Our XAI analysis provides compelling evidence of feature engineering's critical role:

- **Engineered Features Dominate:** Social Activity Score (SHAP importance: 0.303) and Socializing Score (0.213) rank as the two most important features
- **Composite vs Individual:** Engineered features consistently outperform individual behavioral indicators
- **Feature Interaction Validation:** Strong interactions between engineered features (Social Activity Score × Introversion Score: 0.179) confirm their psychological validity

This validates our hypothesis that composite behavioral indicators capture personality traits more effectively than individual features alone, supporting the psychological theory that personality is manifested through complex behavioral patterns rather than isolated actions.

C. Feature Category Importance

Research Question 3: What is the relative importance of social behavior vs. communication patterns vs. lifestyle indicators?

Our SHAP analysis reveals clear hierarchy in feature importance:

- **Social Behavior** (highest importance) - Social Activity Score (0.303), Socializing Score (0.213), Going Outside (0.125)
- **Lifestyle Indicators** (moderate importance) - Time Spent Alone (0.152), Introversion Score (0.158)
- **Communication Patterns** (lower importance) - Stage Fear (0.119), Drained after Socializing (0.018)

This hierarchy suggests that observable social behaviors are the strongest predictors of personality, while communication patterns and lifestyle indicators provide important but secondary signals. This aligns with psychological theory that personality is most clearly expressed through social interaction patterns.

D. XAI Decision-Making Insights

Research Question 4: How do SHAP and LIME techniques reveal the decision-making process of personality prediction models?

Our XAI analysis provides unprecedented insights into model interpretability:

- **Global Interpretability:** SHAP reveals that engineered features dominate decision-making, with clear feature importance rankings
- **Local Interpretability:** LIME provides sample-specific explanations showing how individual feature values contribute to predictions
- **Feature Interactions:** SHAP interaction analysis reveals complex non-linear relationships between personality indicators
- **Error Analysis:** Misclassification patterns show that individuals with mixed personality indicators are harder to classify

The strong agreement between SHAP and traditional feature importance methods (correlation 0.8) validates the reliability of our XAI analysis and confirms that the model's decision-making process aligns with psychological theory.

E. Ethical Considerations and Limitations

Research Question 5: What are the ethical considerations and limitations of automated personality assessment?

Our analysis reveals several critical considerations:

- **Model Limitations:** 2.57% misclassification rate, with higher errors for individuals with mixed personality indicators
- **Overconfidence Issues:** Some misclassified samples had extremely high confidence (up to 99.95%), indicating potential overconfidence in edge cases
- **Feature Bias:** Heavy reliance on social behavior indicators may not capture introverts who are socially active or extroverts who prefer solitude
- **Privacy Concerns:** Behavioral data collection raises privacy and consent issues
- **Reductionism:** Binary classification oversimplifies the complex spectrum of personality traits

These limitations highlight the need for careful deployment, human oversight, and ongoing validation in real-world applications.

VI. Conclusion and Future Work

This research demonstrates the effectiveness of machine learning approaches for personality prediction using behavioral data, achieving exceptional performance with 97.43% accuracy through comprehensive feature engineering, hyperparameter optimization, and ensemble methods. The superior performance of ensemble methods (ROC-AUC: 0.9850) and advanced tree-based models validates the complexity of personality prediction, requiring sophisticated non-linear modeling capabilities to capture intricate behavioral interactions.

Our feature engineering approach proved highly effective, with engineered composite features (Social Activity Score and Socializing Score) dominating the model's decision-making process. The application of explainable AI techniques (SHAP and LIME) provided unprecedented transparency, revealing that social behavior indicators are the strongest predictors of personality, followed by lifestyle indicators and communication patterns. However, our analysis revealed important limitations, including a 2.57% misclassification rate and potential model overconfidence in edge cases, emphasizing that automated personality assessment should be deployed as a complementary tool rather than a replacement for traditional psychological methods.

Future work should focus on advanced deep learning models including transformer-based architectures and attention mechanisms for capturing temporal patterns in behavioral data. Enhanced explainable AI techniques for deep models, such as attention-based interpretability methods and sophisticated SHAP implementations for neural networks, are crucial for maintaining transparency as models become more complex. Additionally, extending beyond binary classification to multi-class personality frameworks and developing real-time assessment systems would provide more nuanced and practical personality assessment capabilities.

VII. Student Contributions

QiaoQiao Han:

- Data exploration and understanding
- Data cleaning and preprocessing
- Feature engineering and selection
- Logistic regression, SVM, Random Forest, and KNN models
- Ensemble methods development
- Results visualization and analysis
- Report writing

Mohammad Faisal Alam:

- Data exploration and understanding
- Data cleaning and preprocessing
- XGBoost, LightGBM, CatBoost, Extra Trees, and AdaBoost models, MLP and TabNet models
- XAI analysis implementation
- SHAP and LIME analysis, adversarial validation
- Research paper writing and final review

As a group, we contributed equally to this project with balanced effort and collaborative work throughout all phases.

References

[1] C. G. Jung, *Psychological Types*, Collected Works of C.G. Jung, Vol. 6, R.F.C. Hull, Trans. Princeton, NJ, USA: Princeton University Press, 1971.

[2] I. B. Myers and M. H. McCaulley, *Manual: A Guide to the Development and Use of the Myers-Briggs Type Indicator*, Palo Alto, CA, USA: Consulting Psychologists Press, 1985.

[3] J. Weizenbaum, "ELIZA—a computer program for the study of natural language communication between man and machine," Communications of the ACM, vol. 9, no. 1, pp. 36–45, 1966.

[4] S. Mahmood et al., "The application of eXplainable artificial intelligence in studying cognition: A scoping review," Ibrain, vol. 10, no. 3, pp. 245–265, 2024.

[5] R. Hussain, J. Ma, R. Khandelwal, J. Oltmanns, and M. Gupta, "Personality Prediction from Life Stories using Language Models," arXiv preprint arXiv:2506.19258, 2025.

[6] J. Zhu, R. Jin, and K. G. Coifman, "Can LLMs Infer Personality from Real World Conversations?" arXiv preprint arXiv:2507.14355, 2025.

[7] F. Xu et al., "Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges," in NLPCC 2019, LNCS 11839, pp. 563–574, 2019.

[8] K. Yang et al., "Towards Interpretable Mental Health Analysis with Large Language Models," arXiv preprint arXiv:2305.17136, 2023.

[9] B. Zhang et al., "Exploring a Gamified Personality Assessment Method through Interaction with Multi-Personality LLM Agents," arXiv preprint arXiv:2507.04005, 2025.

[10] Kaggle, "Predict the Introverts from the Extroverts," Playground Series - Season 5, Episode 7, 2024. [Online]. Available: https://www.kaggle.com/competitions/playground-series-s5e7/data