Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 3, 2026

PERFORMANCE AND UNCERTAINTY EVALUATION IN MACHINE LEARNING MODELS FOR PERSONALITY CLASSIFICATION

Amna Ashraf[1], Muazzam Ali[*2], Laviza Fatima[3], M. U. Hashmi[4], Asifa Ittefaq[5]

[1,3,5]*Department of Basic Sciences, Superior University, Lahore, Pakistan*
[*2,4]*Department of Computer Science, Superior University, Lahore, Pakistan*

[*2]muazzamali@superior.edu.pk

DOI: http://doi.org/10.5281/zenodo.20523316

Keywords
Personality Classification, Support Vector Machines, Uncertainty Quantification, Mental Health Applications, Machine Learning, Psychological Assessment

Article History
Received: 17 February 2026
Accepted: 02 March 2026
Published: 19 March 2026

Corresponding Author: *
Muazzam Ali

Abstract
The classification of personality types is essential in mental health screening, career counseling, and more specific interventions in health care in line with the UNSDG 3 (Good Health and Well-Being). This paper compares machine learning methods for classifying personality types (Introvert, Extrovert, Ambivert) based on a dataset of 20,000 samples and 31 behavioral and psychological variables. Eight models in total were considered as part of robust preprocessing (removing outliers, selecting features, and having a stratified train-test partition) and considering a new uncertainty quantification framework focusing on predictive, aleatoric, and epistemic uncertainties. The Support Vector Machine (SVM) performed best, with accuracy, precision, recall, and F1 score of 0.990786, 0.990794, 0.990786, and 0.990789, respectively, and minimum values of uncertainty (predictive: 0.0248067 and epistemic: 0.00266215). These findings suggest the potential application of SVM in mental health screening, along with individual well-being interventions, in line with the goals of UN Sustainable Development Goal 3. The methodology we have developed helps us address fundamental uncertainties in the data and is a stronger method for reliable personality classification.

1.0 Introduction

Classification of personality types is a very important topic in the fields of psychology, as well as human and computer interactions. The application of classification in personality types extends to career counseling, personalized marketing, and other areas. Classifying individuals into specific psychological types has become a long-standing practice, utilizing the Myers-Briggs Type Indicator (MBTI) and other personality assessment approaches. However, traditional methods of assessment have been met with challenges, including subjectivity, time-consuming administration, and the reliance on self-reported data [1-4]. The current interest in automated methods powered by data is evident in the digital era, as there is growing interest in personality types that can be accurately observed through observable behavioral traits and demographic factors [5,6].

Nevertheless, the classification of personality remains complicated because it has a multi-dimensional form, traits overlap in different classes, and the person is situational. In different contexts, their behavior may change. Such complexity poses a serious problem for

psychological assessment, as well as for the application of machine learning techniques, except in cases where the issues are unambiguous and fit into accepted typologies [7,8]. The problem is that machine learning provides potent tools that can help in solving these challenges [9-11], as it is possible to find hidden patterns in personality-related data that would not be easily detected with traditional forms of analysis. Intricate algorithms can analyze a variety of behavioral and psychological characteristics simultaneously and identify subtle associations that would be overlooked by human analysts [12,13]. The classifier networks, particularly ensemble methods and neural networks, have demonstrated potential in classifying personality types with remarkable accuracy, as evidenced by features such as social interaction preferences, communication styles, and activity patterns [14,15]. Unlike traditional psychometric testing, machine learning models can make progressively more accurate predictions over time as additional data becomes available and can begin to self-adjust to cultural or demographic differences in how personality manifests in individuals. The abilities provide machine learning with significant value in real-time personality analysis, such as in adaptive user interfaces or team compatibility within an organizational context [16-18].

This paper will achieve this by developing and contrasting machine learning models for personality type classification, while maintaining a rigorous evaluation of varying forms of predictive uncertainty. Compared to earlier studies, where accuracy measures were considered primary, our work provides a detailed breakdown of three major uncertainties: predictive (model confidence), aleatoric (data noise), and epistemic (model knowledge gaps). We present a new framework for uncertainty quantification that allows us to determine under what conditions it is possible to rely on the predictions and when human judgment is required. On a dataset of 20,000 samples with 31 behavioral and psychological variables, we test eight classification algorithms, including SVM, neural networks, and ensemble approaches. This is a multifaceted method because it integrates the powerful preprocessing of raw data with detailed measures of uncertainty, providing users with a useful picture of what models can and cannot do in practice. The innovation of this study lies in its ability to not only measure performance but also analyze uncertainty, thereby extending the current state of affairs in assessing the capabilities and limitations of the model in terms of personality assessment.

2.0 Research Methodology

2.1 Dataset Acquisition

The Kaggle-provided dataset in this study included 31 features, such as demographic, behavioral, and psychological characteristics, obtained from 20,000 samples. Personality Type is identified as the target variable, with three categories: Introvert, Extrovert, and Ambivert shown in Figure 1. The data is artificial, yet realistic patterns are exhibited in character gauges. It provides a well-balanced representation through personality types, allowing one to model and compare reliably. This general layout enables a detailed examination of the connections among the different aspects and personality typification.

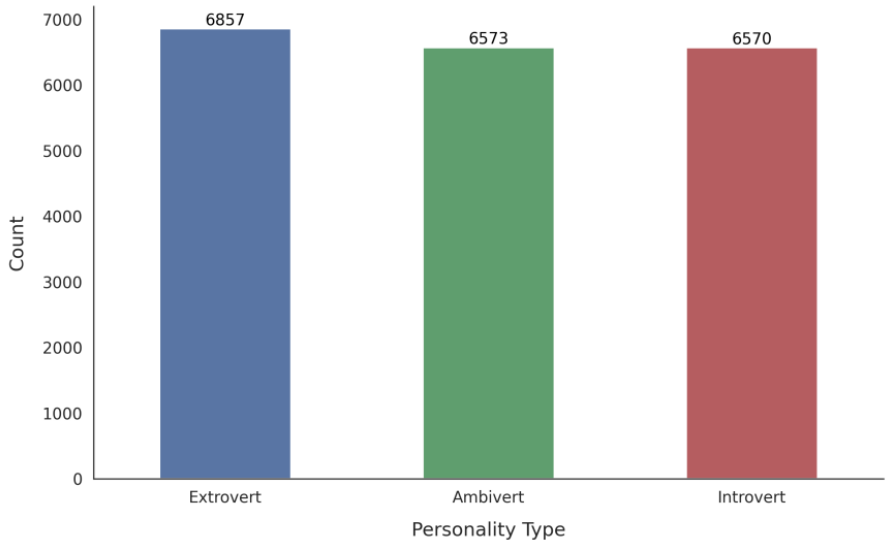

Fig.1. Target Column Distribution Plot

2.2 Preprocessing Techniques

2.2.1 Categorical Variable Encoding

Label encoding converts categories into numbers by assigning each category a unique number. This enables algorithms to process the data, albeit with the risk of implicating false ordinal relationships [19]. Effective on tree-based models, it can be deceiving to linear models, which expect forms of numerical order. The experiment was systematically implemented, ensuring uniformity in the training and test states.

2.2.2 Target Variable Encoding

The target variable, personality_type, was coded into numeric labels (e.g., 0, 1, 2) in a manner compatible with the model. Unlike one-hot encoding, this technique does not increase the dimension, and it does not imply any intrinsic ordering of classes. As personality types are nominal, this method enables the maintenance of interpretability alongside algorithmic needs. The original terms were retained to ensure greater comprehensibility in the assessment.

2.2.3 Outlier Removal using IQR Method

The IQR technique was used to filter outliers by dropping the values that are outliers [20]. This is a harsh method that minimizes skewness as much as possible; however, it may eliminate valid extreme cases. This post-cleaning input (19,209 samples) suggests a fair number of outliers (not excessive, like in noise reduction, but sufficient to warrant consideration).

$$\text{Lower Bound} = Q1 - 1.5 \times IQR \quad (1)$$

$$\text{Upper Bound} = Q3 + 1.5 \times IQR \quad (2)$$

2.2.4 Feature Normalization (Standard Scaling)

Numerical features were standardized to have a mean of zero and a variance of one.

$$z = \frac{x - \mu}{\sigma} \quad (3)$$

This approach avoids the scale bias prevalent in distance-based models (e.g., SVM, KNN) and can accelerate gradient-based optimization [21]. Nonetheless, it assumes that the distributions are Gaussian-like and can destroy any sparse information. The loss of information involved in the tradeoff between better convergence and the likely loss of information was regarded as acceptable.

2.2.5 Feature Selection

While effective for linear relationships, it overlooks non-linear interactions [22]. The reduction from 30+ features enhances efficiency but risks discarding weakly correlated yet useful predictors in complex models, such as Random Forests.

$$\text{Anova F} - \text{Test} = \frac{\text{Between-Group Variance}}{\text{Within-Group Variance}} \quad (4)$$

Table 1: Feature Importance Uncertainty Results (Bootstrapped)

Rank	Feature	Mean Importance	Std Importance	CV Importance
1	party_liking	0.260545	0.004927	0.018912
2	public_speaking_comfort	0.131558	0.009803	0.074512
3	excitement_seeking	0.128135	0.012465	0.097280
4	alone_time_preference	0.124275	0.010652	0.085714
5	talkativeness	0.120754	0.010275	0.085094
6	social_energy	0.109063	0.010230	0.093797
7	reading_habit	0.034921	0.003836	0.109853
8	leadership	0.032801	0.003429	0.104543
9	group_comfort	0.031194	0.002185	0.070041
10	adventurousness	0.026753	0.002512	0.093898

The analysis reveals that party liking is the least changeable and the most powerful predictor of personality type, with a great mean importance (26.05%) and small variability (CV = 1.89%) across bootstrap samples, as shown in Table 1. These findings align with psychological theories, which suggest that extroversion is closely linked to the preference for social engagement. The top six features —public speaking comfort, excitement seeking, and talkativeness—account for 87 percent of the predictive power, indicating that these features constitute a strong personality discriminator. Nonetheless, the rather low importance value of adventurousness (2.68) is not expected according to some theoretical views, which can be linked to the possible shortcomings of the data or conceptual variability with other related constructs such as excitement-seeking. These results demonstrate the importance of both data-driven feature selection and bridging empirical findings with psychological theories, as illustrated in Figure 1.

Tool-wise, the bootstrap methodology serves well to measure the stability of feature importance and discloses serious differences in the predictive precision of traits. Although the social interaction-facing dimensions are very consistent, others, such as reading_habit, have a relatively higher level of uncertainty (CV > 9%), indicating that their predictive potential may hinge on contextual or demographic factors that are not represented in the available model .

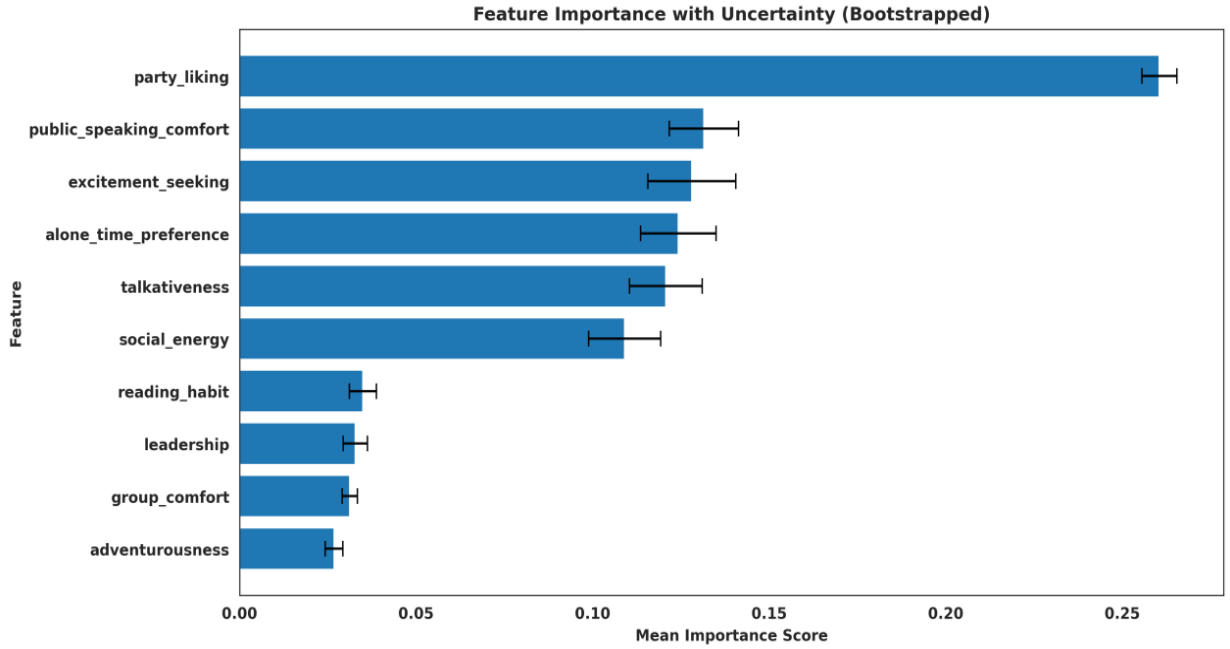

Fig.2 Feature Importance with Bootstrap Results

The high explanatory power of a select few features raises relevant concerns regarding possible redundancy and the selection of these features, as well as the potential for improving the model's parsimony without compromising accuracy through reduction in dimensionality. These findings raise two important tensions in the personality modeling field: data-driven techniques can discover strong predictors of personality variables but may not capture the complex interplay of variables assumed by theoretical models, especially with algorithms such as Random Forest that propose features as independent of each other.

The limitations of the study, such as its use of synthetic data and the independence of features between them, do not allow generalizing these results. Future research needs to confirm these patterns in real datasets and describe alternative models that would be more efficient in terms of modeling psychological construct interactions. However, the positive results of primary social engagement characteristics provide excellent arguments for their central position in personality measurement, and the inconsistency of secondary traits strengthens the selective manifestation of personality in different circumstances. These understandings should inform the continuing debate on the balance between theory-guided exploration and data-driven computational personality study.

2.3 Train-Test Split

A stratified split preserved class proportions in both the training (70%) and test (30%) sets, ensuring unbiased evaluation—fixed randomization (random_state=42) guaranteed reproducibility. Although ideal for balanced datasets, stratification alone doesn't address severe class imbalances, which may require additional techniques, such as resampling, for optimal performance.

2.4 Used Models

2.4.1. Decision Tree Classifier

The max_depth hyperparameter of the Decision Tree model was set to 5 to avoid overfitting without compromising interpretability. The reproducibility of splits was ensured by setting random_state=42. The model can control bias and variance by imposing depth restrictions on trees, but this tradeoff may compromise some of its prediction capability for complex patterns. Splitting was based on the Gini impurity criterion (default), which is efficient to measure

the purity of nodes, although it can be biased towards high-cardinality features. This model serves as both a baseline and a component in ensemble methods [23].

2.4.2. Support Vector Machine (SVM)

SVM deployed RBF kernel (kernel='rbf'), C=1.0 (regularization), and gamma='scale' (auto-adjusted coefficient of kernel). The RBF kernel represents a non-linear relationship, whereas C governs overfitting, as higher values will fit the training data closely and tend to undermine generalization. The parameter probability=True enabled the estimation of probabilities in the stacking. Though powerful, SVMs' computational cost scales poorly with large datasets, and their performance heavily depends on proper kernel and hyperparameter tuning [24].

2.4.3. K-Nearest Neighbors (KNN)

The parameters of KNN were n neighbors=5, and the distance, which was calculated by default, was Euclidean distance. It is a local, instance-based model (i.e., based on a local data structure), and its disadvantages, including memory consumption and slow prediction with large datasets, are due to the use of a local data structure. k = 5 is a compromise between noise sensitivity and underfitting, but ideally, k should be determined through cross-validation. Feature scaling (done earlier) is critical for KNN, as distance metrics are scale-dependent [25].

2.4.4. Random Forest Classifier

The 200-tree ensemble (n estimators = 200) was trained to be robust in feature interactions by setting max_depth to unlimited and min_samples_split to 5, thereby preventing overfitting. The specified randomness was fixed to prevent randomness and ensure reproducibility. The feature subsampling in Random Forest (default: max_features='sqrt') increases the diversity between trees to enhance generalization. However, its black-box nature limits interpretability, and training time grows linearly with the number of trees [26].

2.4.5. XGBoost Classifier

To balance speed and performance, XGBoost was optimized by setting n_estimators to 200, max_depth to 5, and learning_rate to 0.1. The slim trees (max_depth) prevent overfitting, and the medium learning rate produces stable convergence. The eval_metric: = mlogloss was optimized using multiclass logarithmic loss. The fact that XGBoost implements regularization and handles missing values internally, as well as its sensitivity to well-tuned hyperparameters, makes it robust [27]; however, this aspect was not thoroughly tested in this work.

2.4.6. Bagging Classifier

A 50-bagging ensemble of 5-depth Decision Trees (n_estimators=50) was used to decrease variance with the help of bootstrap aggregation. The seeded sampling was ensured by random_state=42. Bagging provides stability when compared to a single tree; however, when the base models are already good (e.g., when compared to Random Forest), the gain is reduced. The choice of Decision Trees as base estimators prioritizes simplicity over potential gains from more complex weak learners [28].

2.4.7. Artificial Neural Network (ANN)

MLPClassifier employed one hidden layer of 100 neurons (hidden_layer_sizes=(100,)), and the max_iter=500 iterations. Early stopping, with True, stopped the training when the validation scores were stagnating and thereby avoided overfitting. It kept the default activation, ReLU activation, and Adam optimizer. The limitations of ANNs in modeling complex patterns may be present here due to their simple architecture and the absence of hyperparameter optimization (e.g., learning rate, number of layers). Scalability to larger datasets is also constrained by computational cost [29].

2.4.8. Stacking Classifier

A meta-model whose list of estimators included predictions based on a Decision Tree (max_depth=5), SVM (all the same parameters as the one above), and a final estimator as a

Random Forest (n_estimators=100). The stacking was done using a 5-fold cross-validation (cv=5) that gave strong meta-features. The idea of stacking is to leverage the complementary abilities of base models [30], but the success of this approach relies on the heterogeneity of those model sets. In this case, both learning linear (SVM) and non-linear (Decision Tree) models were chosen [31]. However, the added complexity may not always justify marginal accuracy gains over standalone ensembles, such as Random Forest [32].

2.5 Data Visualizations

Figure 3 illustrates that, due to its well-structured and visually intuitive form, a critical comparison of five psychological traits according to personality types — Extrovert, Ambivert, and Introvert — can be made using box plots. Through the Social Energy plot, it is clear that extroverts score the highest, indicating that they are very active in societal matters. Conversely, the lowest score obtained suggests that introverts employ a minimum stimulation strategy. By contrast, the Alone Time Preference diagram indicates that introverts are most interested in solitude, followed closely by ambiverts, while extroverts do not tend to concern themselves with solitude much. The same pattern can be observed in talkativeness, where the talkative individuals, namely the extroverts, are the highly verbally expressive group, and introverts, the least verbally expressive, due to their naturally withdrawn nature. Party-loving extroverts take the lead in this regard, with their social event cheering, whereas introverts express their mild interests, and ambiverts remain moderate. Deep Reflection, however, presents an opposite trend, with introverts being the most introspective, extroverts the least, and ambiverts still lying in between. In all characteristics, ambiverts will always be at the midpoint between the two extremes, therefore proving their adaptive behavior. The clear divergence of medians and interquartile ranges substantiates the discriminating ability of the chosen traits. These perceptual patterns have a close connection with the established psychological systems of personality behavior. The individual variability may be demonstrated by outliers in each given trait, whereas the high consistency of all traits supports the reliability of the collected data.

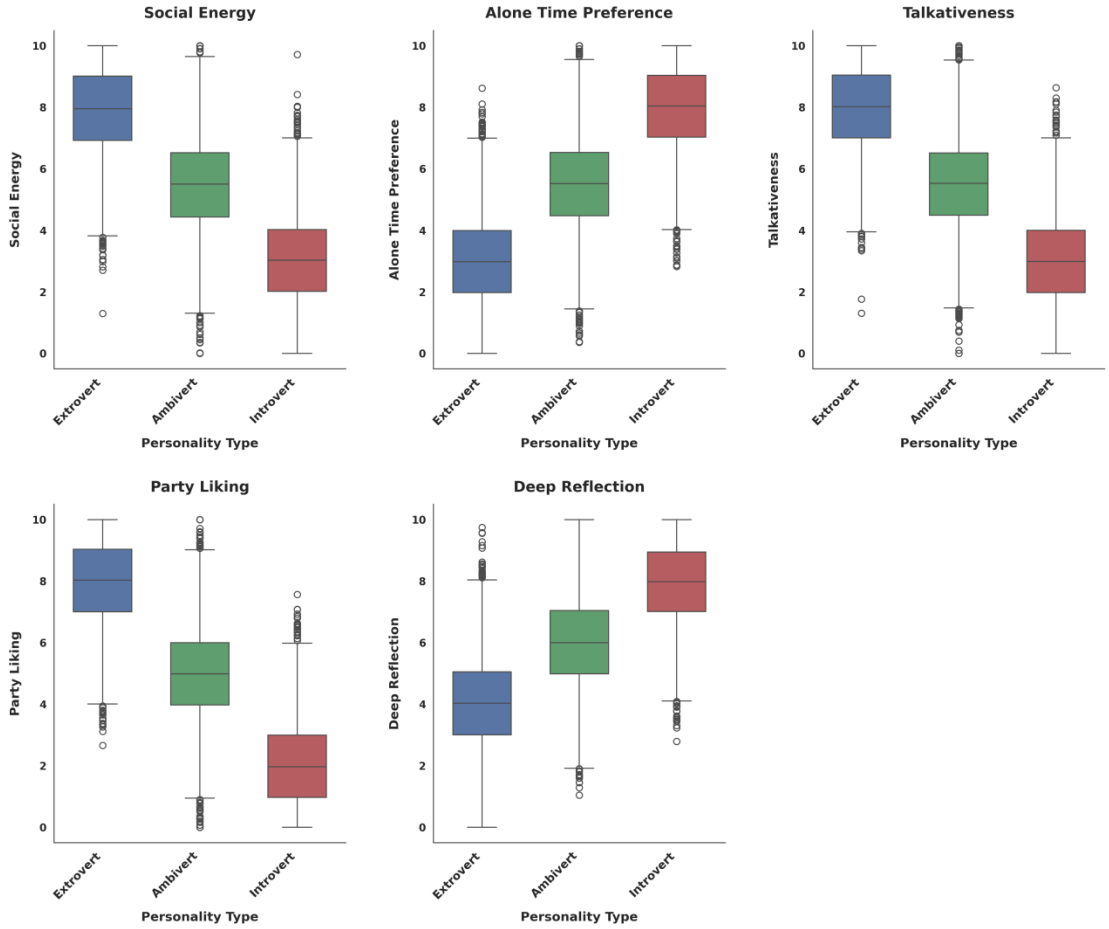

Fig.3. Box Plot of Psychological Traits

The violin plots shown in Figure 4 are visualizations of the distribution and density of 5 psychological traits related to personality types, using smoothed kernel density estimates. The extroverts consistently have stronger scores on Social Energy, Talkativeness, and Party Liking, with concentrated upward-stretched distributions serving as evidence of high group unanimity on these variables. On the other hand, introverts have a high Alone Time Preference and Deep Reflection, which supports their reflective and withdrawn nature at higher levels. Ambiverts exhibit wider, evenly balanced distributions across all personality traits, considering they possess a flexible type of personality. The first aspect is that the broader shapes in some sections give more information than the box plots to understand where the concentration of responses is. High clustering on the extremes, particularly among extroverts and introverts, indicates a high degree of behavioral polarization. Notably, the plots do not show any overlap between the groups, although they exhibit a characteristic distribution pattern. Overall, the figure enhances the understanding of trait variation and validates theoretical distinctions between types of personality.

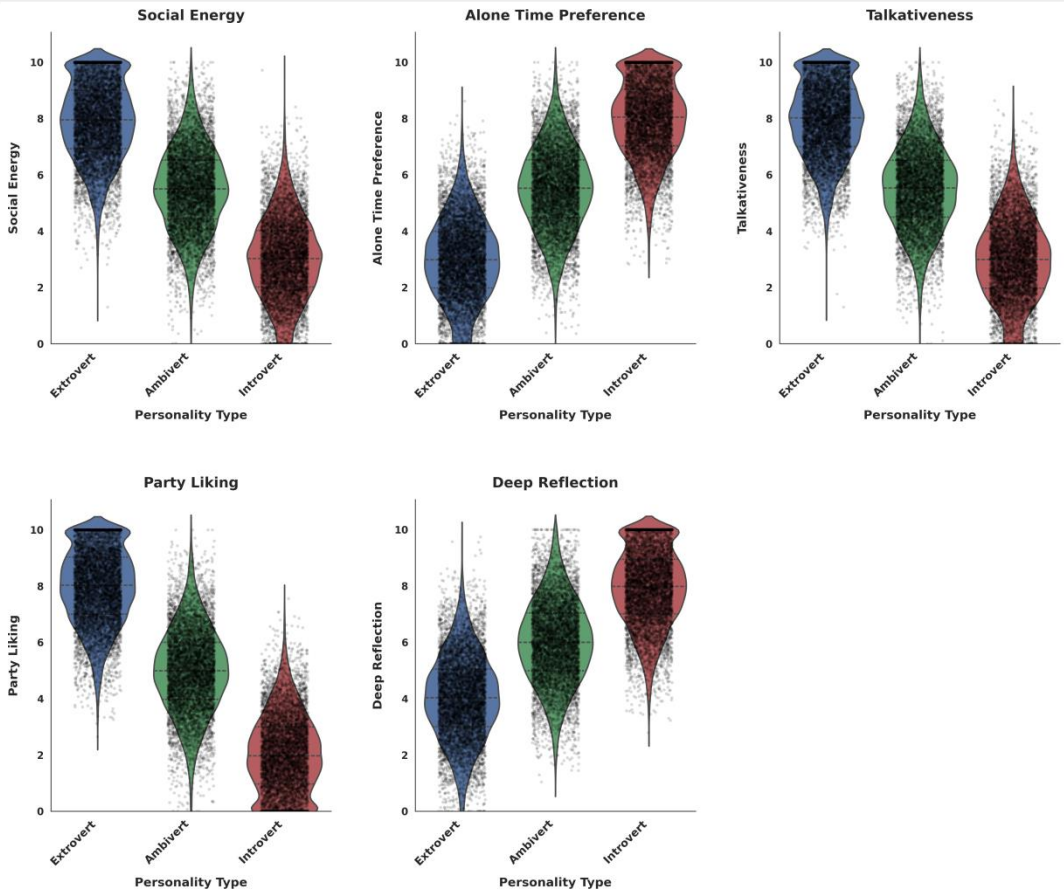

Fig.4. Violin Plot of Personality Traits

2.6 Uncertainties

2.6.1 Predictive Uncertainty

Predictive uncertainty estimates the level of uncertainty in model predictions at the instance level, specifically the confidence level of the model in making classifications [33]. This is calculated as the entropy of the probabilities of the predicted classes, with higher entropy signifying high uncertainty. Its mathematical equation is:

$$H(p) = -\sum_i p_i \log p_i \quad (5)$$

In which p_i is the predicted probability in each class. In the study, predictive uncertainty is useful in identifying situations where the model's predictions are more uncertain, especially in personality measures, where ambiguous cases require the attention of human experts. We computed it for every test case and averaged it over the dataset to allow for comparison of uncertainty levels among different models.

2.6.2 Aleatoric Uncertainty

Aleatoric uncertainty is the fundamental noise, or irreducible uncertainty in the data, which reflects simply measurement error or actual ambiguity in personality aspects [34]. We also calculated it by generating bootstrap samples of predictions and computing the mean-averaged entropy of the mean predicted probabilities.

$$\text{Aleatoric} = E[H(\bar{y})] \quad (6)$$

where $\bar{y}$ is the mean prediction over bootstrap samples and H is the entropy. This ambiguity is reflective, as it highlights inherent drawbacks in the data that cannot be mitigated through improved modeling. We found that the values of aleatoric uncertainty were high in all models we examined, indicating that the possible difficulties in personality type classification may require improvement in the methods of data collection.

2.6.3 Epistemic Uncertainty

Dutch: Epistemic uncertainty quantifies uncertainty in the model's parameters due to the limited training data, which is what the model does not know [35]. We have estimated it as an averaged standard deviation in the prediction of bootstrap models:

$$\text{Epistemic} = \text{E}[(\text{y})] \quad (7)$$

With ensemble techniques such as Random Forest, we ran the available estimators, and with other models, we bootstrapped ensembles [36]. This uncertainty is significant as it diminishes with the increase in the training dataset and signals in which situations the model can be improved. Our results indicate that ensemble methods have higher epistemic uncertainty, which is inherent to ensemble methods, particularly in comparison to simple models such as SVM, which are complex and data-intensive. The mathematical model will use the determination of a variance in predicting:

$$\sigma_d^2 = E[(y_d - \bar{y}_d)^2] \quad (8)$$

2.7 Evaluation Metrics

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (9)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (10)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (11)$$

$$\text{F1} - \text{Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision+Recall}} \quad (12)$$

$$\text{CV Accuracy} = \frac{1}{\kappa}\sum_{\text{i}=1}^{\kappa} \text{Accuracy}_{\text{i}} \quad (13)$$

3.0 Results and Discussions

The performance comparison of eight classification models —SVM, ANN, Stacking, KNN, XGBoost, Random Forest, Bagging, and Decision Tree —was conducted using five evaluation metrics: Accuracy, Precision, Recall, F1 Score, and Mean Cross-Validation (CV) Accuracy. The results are shown in Table 2.

Table 2. Performance Evaluation Results of Models

Model	Accuracy	Precision	Recall	F1 Score	Mean CV Accuracy
SVM	0.990786	0.990794	0.990786	0.990789	0.989716
ANN	0.990438	0.990434	0.990438	0.990435	0.988822
XGBoost	0.988526	0.988529	0.988526	0.988527	0.986735
KNN	0.988004	0.988009	0.988004	0.988006	0.98763
Random Forest	0.987135	0.987148	0.987135	0.987139	0.986139
Bagging	0.953408	0.953792	0.953408	0.953515	0.948209
Decision Tree	0.903686	0.904052	0.903686	0.903761	0.900514

3.1 Results of Evaluation Metrics

The findings of the conducted experiment indicate that SVM and ANN are the best models, as these models demonstrated the highest accuracy (0.990786 and 0.990438, respectively), with relatively perfect precision, recall, and F1-scores (all around 0.990). Its good performance can be explained by the fact that it handles feature scaling well and can generate non-linear insights in data, which are very important features of model complex psychological model. The other methods, such as XGBoost (0.988526), KNN (0.988004), and Random Forest (0.987135), also performed fairly well, indicating the strength of ensemble and distance-based techniques. Instead, Bagging (0.953408) and Decision Tree (0.903686) came far behind, indicating that simple models fail to account for high-dimensional interactions between features unless they are boosted or averaged.

The precision, recall, and F1-scores of the models also exhibited a similar trend, with SVM and ANN showing balanced classification rates across all types of personalities, indicating that there was no bias in the predictions. Bagging and Decision Tree, however, yielded lower precision results of 0.953792 and 0.904052, respectively, indicating a greater risk of misinterpreting border cases, such as the difference between Ambivert and Extrovert. The cross-validation observations also

supported the consistency of high-scoring models, whereby SVM (0.989716) and ANN (0.988822) showed less variance. At the same time, Bagging (0.948209) and Decision Tree (0.900514) were more affected by the data partitioning, demonstrating their disadvantage in terms of generalization (See Figure 5).

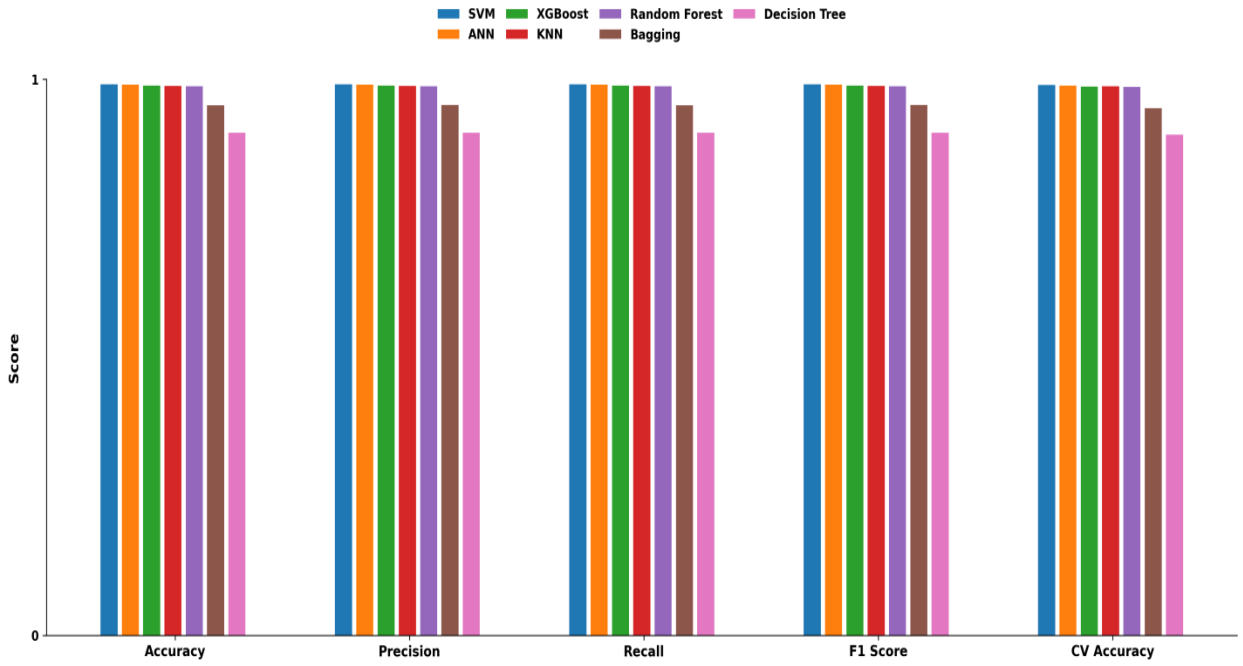

Fig.5. Comparison of Performance Evaluation of Models

The results indicate that the performance of kernel-based methods (SVM) and neural networks (ANN) in classifying personality traits is more suitable because these models can implement complex decision boundaries. Although ensemble methods, such as XGBoost and Random Forest, continue to show competitive results compared to other methods, their slight disadvantage in the performance spectrum highlights the possibility of hyperparameter optimization. Conversely, other simpler models, such as Bagging and Decision Trees, will require either more complex architectures or alternative preprocessing to enhance reliability. The consistency of the high values of precision, recall, and F1-scores of the best models indicates that the preprocessing pipeline used is efficient. In contrast, the range of different results obtained from models using advanced and more basic methods suggests the importance of attention to model choice in personality computation. Future research would be interesting to consider hybrid models or real-world data to confirm these findings further.

The given confusion matrices (Figure 6) present comparative results of eight machine learning models considered: Decision Tree, SVM, KNN, Random Forest, XGBoost, Bagging, ANN, and Stacking, regarding the process of classifying personality predictors (Ambi, Extro, Intro). On the whole, models such as SVM, KNN, Random Forest, XGBoost, ANN, and Stacking have promising accuracy, as their diagonal (i.e., the accurate predictions) is high, and the instances of misclassification are low. As an example, SVM and KNN demonstrate a close-to-perfect classification of Extro and Intro, while Ambi is somewhat confused with the rest. On the contrary, Decision Tree and Bagging have greater misclassification rates, especially for Ambi, implying lower performance. The findings indicate that ensemble techniques (e.g., Random

Forest, XGBoost) and neural networks (ANN) outperform simpler approaches, such as Decision Trees. This discussion highlights the theme of the need to choose a model of personality trait classification that can be more reliable with the implementation of advanced approaches.

3.2 Uncertainty Quantification

Table 3 shows the uncertainty results of different models.

Table. 3 Uncertainty Evaluations of Different Models

Model	Prediction Uncertainty	Aleatoric Uncertainty	Epistemic Uncertainty
SVM	0.0248067	1.08853	0.00266215
ANN	0.037218	1.08873	0.0102408
XGBoost	0.0256185	1.08836	0.00544322
KNN	0.0253168	1.08838	0.00808957
Random Forest	0.162529	1.08974	0.095754
Bagging	0.34391	1.0913	0.0930444
Decision Tree	0.25952	1.0905	0.0941059

The Bagging Classifier had the largest variance in prediction (0.34391), while the Decision Tree and Random Forest had variances of 0.25952 and 0.162529, respectively. This implies that although ensemble techniques tend to produce more accurate results, they can also be more likely to generate higher levels of uncertainty in probabilistic predictions. However, in comparison, SVM revealed significantly lower uncertainty in the predictions (0.0248067), indicating that SVM is more confident in its predictions; yet, no connection with accuracy was shown in Figure 7. The ANN (0.037218) and XGBoost (0.0256185) exhibited moderate levels of uncertainty, indicating that they are neither overconfident nor underconfident about their predictions, but rather are flexible with their predictions.

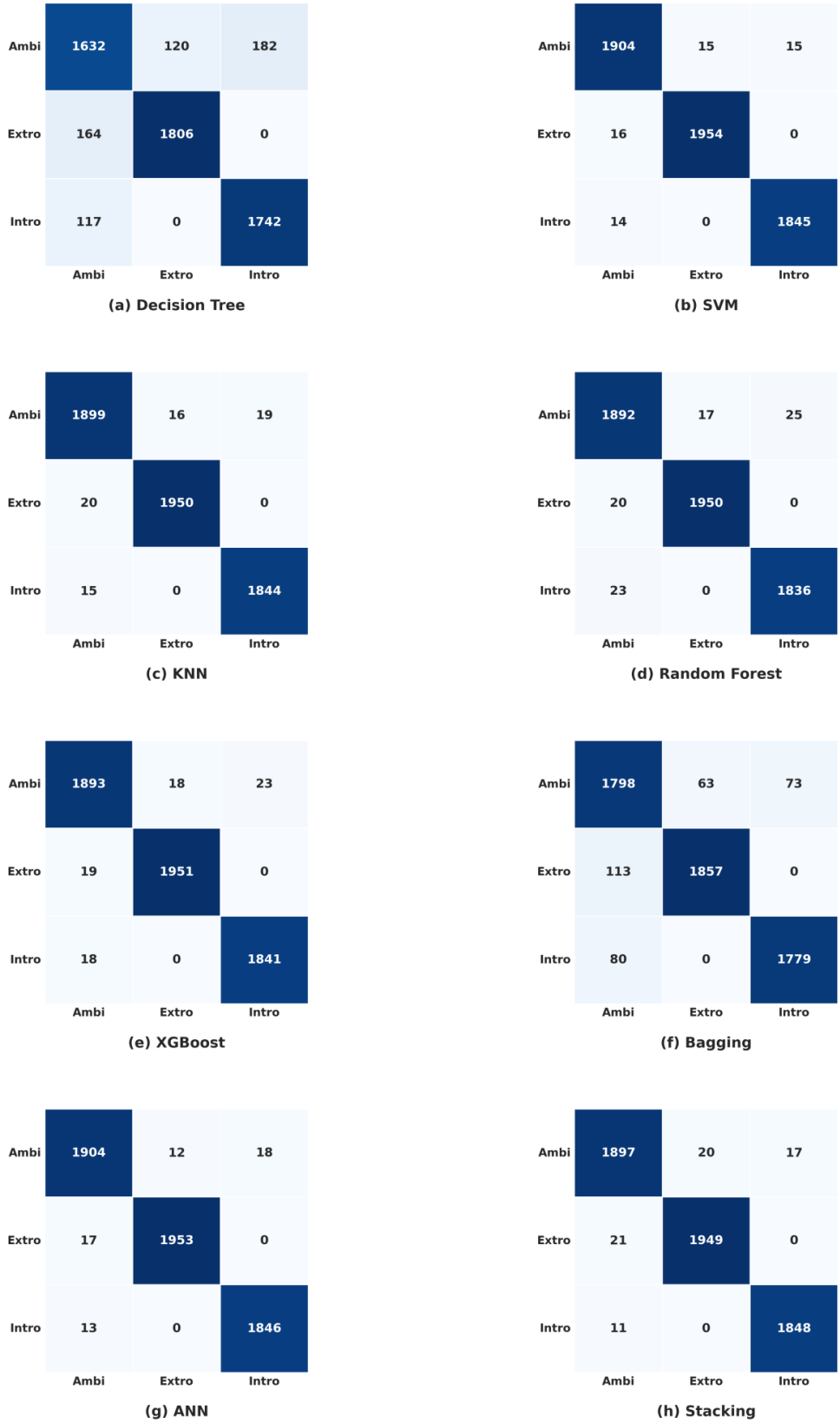

Fig.6. Confusion Matrix Results of Models

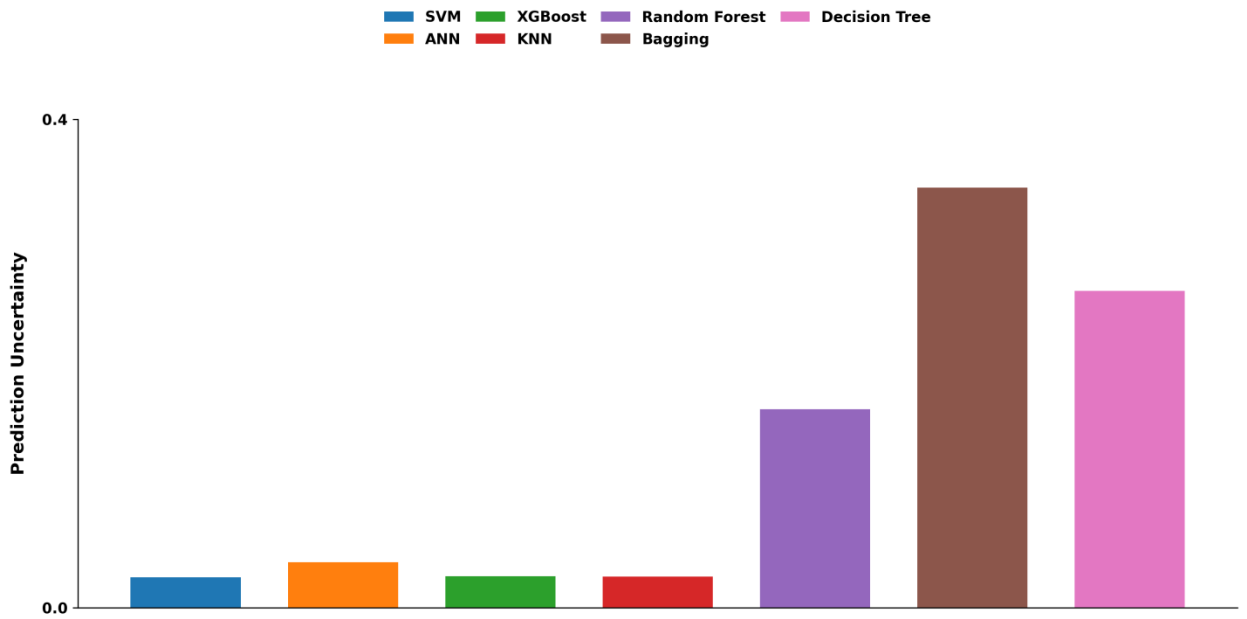

Fig.7 Prediction Uncertainty Comparison across Models

Aleatoric uncertainty was also high on all models (ranging from -1.08836 to 1.0913) shown in Figure 8, which means that there was noise in the data or this uncertainty could not be removed. This implies that some inherent factors contribute to the inconsistency in the classification of different personality types, which could be due to measurement deficiencies or the actual vagueness in the classification of personality. The epistemic uncertainty differed significantly between models (0.00266215-0.095754), with higher values recorded among the Random Forest, Bagging, and Decision Tree models. This pattern implies that ensemble methods maintain more model uncertainty, which may be due to the more complex decision boundaries resulting from the heterogeneity of the comprised models.

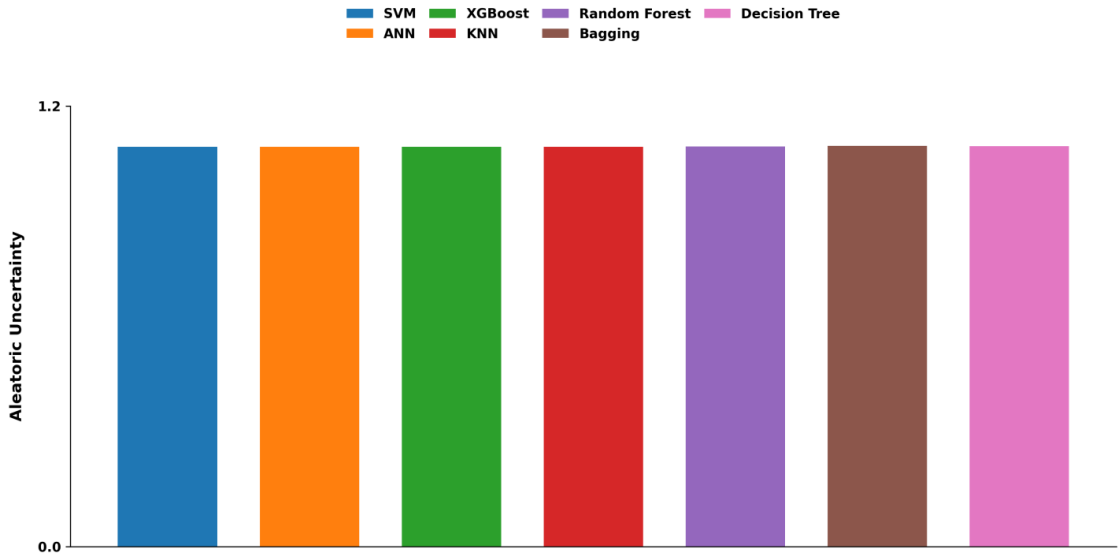

Fig.8. Aleatoric Uncertainty across Models

The epistemic uncertainty of SVM is by far the lowest (0.00266215), indicating that SVM tends to a very stable solution, albeit at the expense of possibly less flexible parameters. The moderate ANN values of uncertainty (0.037218 prediction, 0.0102408 epistemic) shown in Figure 9 indicate that the ANN does not overcommit on prediction and does not have an excess of epistemic uncertainty related to its model. Epistemic uncertainty in ensemble methods (Random Forest, Bagging) was the largest, and this can be explained by the design wisdom behind them, which aims to encompass many different weak learners as a mechanism to promote epistemic uncertainty, thereby resulting in models that generalize well.

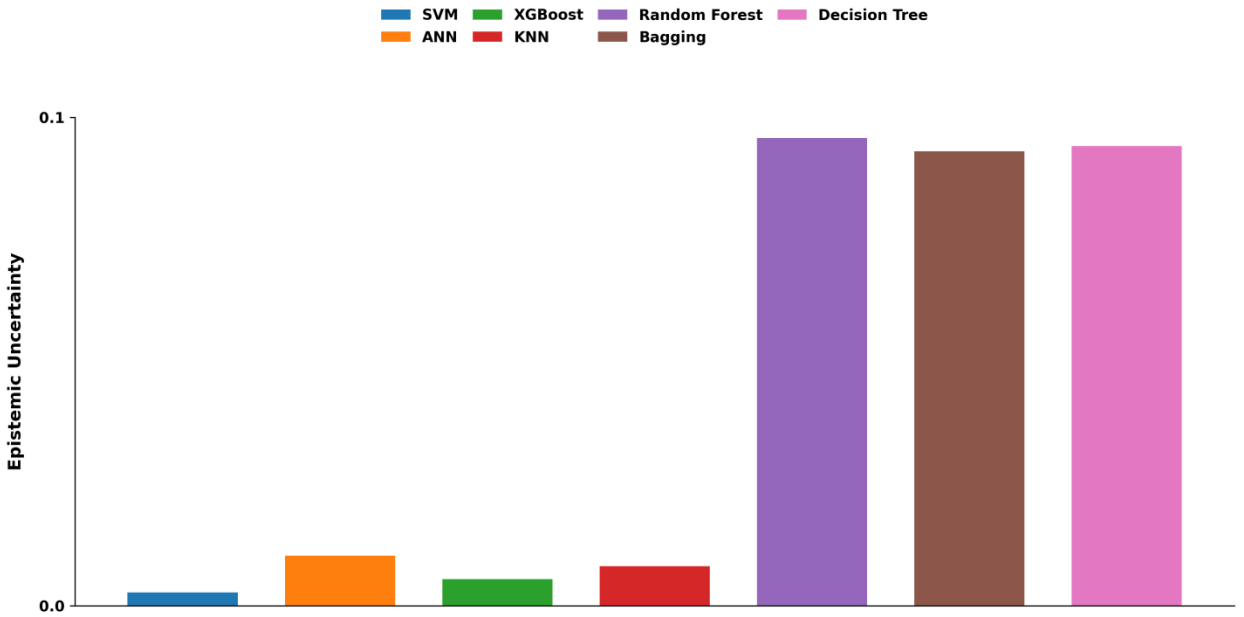

Fig.9. Epistemic Uncertainty Comparison Across Models

Through this study, we successfully demonstrated the potential application of machine learning models, specifically Support Vector Machines (SVM) and Artificial Neural Networks (ANN), in the classification of personality types based on behavioral and psychological information. We were able to extract performance evaluations as well as reliability values of the predictions due to our multifaceted strategy incorporates not only solid preprocessing, uncertainty measurements, and performance estimations but also fulfills the generalization of models. The findings demonstrate that SVM, with its high accuracy and low levels of uncertainty, proves to be a promising tool for application in practice in the field of mental health screening and well-being interventions. The ensemble techniques, such as Random Forest and XGBoost, were competitive; however, their uncertainty value was larger than that of other methods, and their accuracy was slightly lower. Therefore, ensemble methods are not recommended when precision and confidence in predictions are crucial. The introduction of predictive, aleatoric, and epistemic measures of uncertainty has provided more in-depth information about the pros and cons of each model, allowing for a more informed decision in the area of psychological assessment.

4.0 Conclusion

In conclusion, results suggest that SVM performed the best of all the considered models, with the accuracy of 0.990786 and an exceptionally low value of uncertainty (predictive uncertainty: 0.0248067, epistemic uncertainty: 0.00266215), which means not only that the considered model has high predictive power but

also that the model is reliable. Whereas the aleatoric uncertainty, ranging from 1.08853 to 1.0913, was consistently high in all models, indicating intrinsic noise in the personality data, the high accuracy and low uncertainty of the SVM model suggested that it was highly suitable for highly sensitive applications in mental health and well-being domains. Artificial Neural Network also did quite well, reaching an accuracy of 0.990438, but with greater uncertainty measures. These results indicate that although the performance of ensembles such as Random Forest (accuracy: 0.987135) and XGBoost (accuracy: 0.988526) is good, it may be beneficial to apply simpler models, such as SVM, when both precision and prediction confidence are crucial. The uncertainty-minded strategy of the study provides meaningful guidance for the development of trustworthy artificial intelligence in the field of psychological evaluation, particularly in applications that support the objectives of UNSDG 3, which include good health and well-being. Further investigation is necessary to confirm all the results with actual clinical data and consider hybrid solutions that combine the best features of all the models, thereby reducing the uncertainties associated with each of them.

Acknowledgment

We would also like to thank very much everyone who helped us with this research. We also give our sincere gratitude to the Kaggle platform, which made the dataset utilized in this paper freely available. This is highly appreciated by the Department of Basic Sciences, Superior University, Lahore, and the Department of Computer Science, Superior University, Lahore, and it was made possible with their constant support. We also acknowledge the constructive advice of our fellow scholars at the Department of Information Technology Development, Manchester Metropolitan University, UK, whose guidance has significantly enhanced the quality of this work. Lastly, we would like to extend our gratitude to the anonymous reviewers.

Ethical Approval

In the study, publicly accessible and synthetic datasets were utilized, and no human subject participation was involved. Therefore, there was no need to obtain ethical approval. Nevertheless, we were guided by the ethical research guidelines when managing and processing the data, and the privacy and integrity of the information were preserved throughout the study.

Conflict of Interest

The authors have no conflict of interest.

REFERENCES

Ishkov A. Myers-Briggs typology from the perspective of classification of natural specializations. Edelweiss Applied Science and Technology. 2025;9(5):979-92.

Liying W, Sheibani S. The Relationship between the Myers-Briggs Type Indicator (MBTI) Types and Psychological Well-being among College Students in China. Journal of Ecohumanism. 2024 Oct 28;3(7):3611-9.

Ma X. A Comparison of MBTI and Big Five Personality. Interdisciplinary Humanities and Communication Studies. 2025 Jun 26;1(3).

Bhanushali J, Goswami N, Bhatt N, Spoorthy V. Personality Prediction Using Myers-Briggs Kind Indicator and Machine Learning Approaches. In Congress on Intelligent Systems 2025 (pp. 473-484). Springer, Singapore.

Hinds J, Joinson AN. Digital data and personality: A systematic review and meta-analysis of human perception and computer prediction. Psychological Bulletin. 2024 Jun;150(6):727.

Bhadane SN, Verma P. Review of Machine Learning and Deep Learning algorithms for Personality traits classification. In 2024 2nd DMIHER International Conference on Artificial Intelligence in Healthcare, Education and Industry (IDICAIEI) 2024 Nov 29 (pp. 1-6). IEEE.

Dobrosovestnova A, Reinboth T, Weiss A. Towards an integrative framework for robot personality research. ACM Transactions on Human-Robot Interaction. 2024 Feb 24;13(1):1-22

Chen X, Chen Y, Yin G. Exploring the motivations behind behavior: A theory-driven deep-learning framework for cyberviolence behavior detection. Decision Support Systems. 2025 Jan 28:114409.

Sánchez-Fernández P, Ruiz LG, Jiménez MD. Application of classical and advanced machine learning models to predict personality on social media. Expert Systems with Applications. 2023 Apr 15;216:119498.

Alslaity A, Orji R. Machine learning techniques for emotion detection and sentiment analysis: current state, challenges, and future directions. Behaviour & Information Technology. 2024 Jan 2;43(1):139-64.

Samsuryadi, Kurniawan R, Supardi J, Sukemi, Mohamad FS. A framework for determining the big five personality traits using machine learning classification through graphology. Journal of Electrical and Computer Engineering. 2023;2023(1):1249004.

Gigerenzer, G. Psychological AI: Designing algorithms informed by human psychology. Perspectives on Psychological Science. 2024 Sep;19(5):839-48.

Kleinberg J, Ludwig J, Mullainathan S, Raghavan M. The inversion problem: Why algorithms should infer mental state and not just predict behavior. Perspectives on Psychological Science. 2024 Sep;19(5):827-38.

Naz A, Khan HU, Bukhari A, Alshemaimri B, Daud A, Ramzan M. Machine and deep learning for personality traits detection: a comprehensive survey and open research challenges. Artificial Intelligence Review. 2025 Aug;58(8):1-57.

Alsini R, Naz A, Khan HU, Bukhari A, Daud A, Ramzan M. Using deep learning and word embeddings for predicting human agreeableness behavior. Scientific Reports. 2024 Dec 2;14(1):29875.

Khamaj A, Ali AM. Adapting user experience with reinforcement learning: Personalizing interfaces based on user behavior analysis in real-time. Alexandria Engineering Journal. 2024 May 1;95:164-73.

Vargas EP, Carrasco-Ribelles LA, Marin-Morales J, Molina CA, Raya MA. Feasibility of virtual reality and machine learning to assess personality traits in an organizational environment. Frontiers in Psychology. 2024 Jul 24;15:1342018.

Semwal R, Tripathi N, Rana A, Pandey UK, Parihar S, Bairwa MK. AI-Driven Insights: Enhancing Personality Type Prediction with Advanced Machine Learning Algorithms. In 2024 7th International Conference on Contemporary Computing and Informatics (IC3I) 2024 Sep 18 (Vol. 7, pp. 815-822). IEEE.

Bolikulov F, Nasimov R, Rashidov A, Akhmedov F, Young-Im C. Effective methods of categorical data encoding for artificial intelligence algorithms. Mathematics. 2024;12(16):2553.

Alves F, Souza EG, Sobjak R, Bazzi CL, Hachisuca AM, Mercante E. Data processing to remove outliers and inliers: A systematic literature study. Revista Brasileira de Engenharia Agrícola e Ambiental. 2024 Jul 22;28(9):e278672.

Hasan R, Biswas B, Samiun M, Saleh MA, Prabha M, Akter J, Joya FH, Abdullah M. Enhancing malware detection with feature selection and scaling techniques using machine learning models. Scientific Reports. 2025 Mar 17;15(1):9122.

Cheng X. A Comprehensive Study of Feature Selection Techniques in Machine Learning Models. Insights in Computer, Signals and Systems. 2024 Nov 25;1(1):10-70088.

Zhou ZH. Ensemble methods: foundations and algorithms. CRC press; 2025 Feb 15.

Ali YA, Awwad EM, Al-Razgan M, Maarouf A. Hyperparameter search for machine learning algorithms to optimize computational complexity. Processes. 2023 Jan 21;11(2):349.

Chai BX, Eisenbart B, Nikzad M, Fox B, Blythe A, Bwar KH, Wang J, Du Y, Shevtsov S. Application of KNN and ANN metamodeling for RTM filling process prediction. Materials. 2023 Sep 7;16(18):6115.

Stempfle L, Matsson A, Mwai N, Johansson FD. Prediction Models That Learn to Avoid Missing Values. arXiv preprint arXiv:2505.03393. 2025 May 6.

Hassija V, Chamola V, Mahapatra A, Singal A, Goel D, Huang K, Scardapane S, Spinelli I, Mahmud M, Hussain A. Interpreting black-box models: a review on explainable artificial intelligence. Cognitive Computation. 2024 Jan;16(1):45-74.

Zhang S, Chen X, Ran X, Li Z, Cao W. Prioritizing causation in decision trees: A framework for interpretable modeling. Engineering Applications of Artificial Intelligence. 2024 Jul 1;133:108224.

Muennighoff N, Rush A, Barak B, Le Scao T, Tazi N, Piktus A, Pyysalo S, Wolf T, Raffel CA. Scaling data-constrained language models. Advances in Neural Information Processing Systems. 2023 Dec 15;36:50358-76.

Roy RK, Bhavsar A, Chandla A, Dutt V. PRAK-XGBoost: Interpretable Two-Level Stacking for Multiclass Ayurvedic Personality Classification from Psychometric Data. Practice. 2025;4:5.

Shaikh TA, Rasool T, Verma P, Mir WA. A fundamental overview of ensemble deep learning models and applications: systematic literature and state of the art. Annals of Operations Research. 2024 Dec 24:1-77.

Fife DA, D'Onofrio J. Common, uncommon, and novel applications of random forest in psychological research. Behavior Research Methods. 2023 Aug;55(5):2447-66.

Tyralis H, Papacharalampous G. A review of predictive uncertainty estimation with machine learning. Artificial Intelligence Review. 2024 Mar 18;57(4):94.

Chiaburu T, Haußer F, Bießmann F. Uncertainty in XAI: Human perception and modeling approaches. Machine Learning and Knowledge Extraction. 2024 May 27;6(2):1170-92.

Wang H, Ji Q. Epistemic Uncertainty Quantification For Pre-trained Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2024 (pp. 11052-11061).

[36] Cheng S, Quilodrán-Casas C, Ouala S, Farchi A, Liu C, Tandeo P, Fablet R, Lucor D, Iooss B, Brajard J, Xiao D. Machine learning with data assimilation and uncertainty quantification for dynamical systems: a review. IEEE/CAA Journal of Automatica Sinica. 2023 May 31;10(6):1361-87.