Nile University
جامعة النيل

CBIO 313

Alzheimer's disease stage prediction Using multimodal machine learning

Name: Ashraqat Mohamed Abdelhamid

ID: 221000836

Under The supervision of: Dr. Muhammed Elsayeh

TA: Malak Abdel Monsef

Abstract

Alzheimer’s disease (AD) is a progressive neurodegenerative disorder characterized by cognitive decline, requiring accurate early-stage classification for effective intervention. This study develops a machine learning pipeline to classify AD stages: Cognitively Unimpaired (CU), Mild Cognitive Impairment (MCI), and AD, using multimodal data comprising clinical assessments and gene expression profiles sourced from the Alzheimer's Disease Neuroimaging Initiative (ADNI). After comprehensive data preprocessing, including feature reduction via Principal Component Analysis and data augmentation with SMOTE, various classification algorithms were evaluated. Ensemble methods, particularly the Stacking Classifier combining Logistic Regression, Random Forest, and XGBoost, demonstrated superior performance, achieving a balanced accuracy of 93% and ROC-AUC close to 0.99 across classes. The final model’s robustness and generalization were confirmed through detailed confusion matrix and ROC curve analyses. The deployed classification system was implemented using Streamlit, providing an accessible interface for real-time AD stage prediction. These results highlight the potential of ensemble machine learning approaches in supporting clinical diagnosis and personalized management of Alzheimer’s disease.

1.0 Introduction

Alzheimer’s disease (AD) is a progressive neurodegenerative disorder and the most common cause of dementia worldwide. It affects memory, cognition, and behavior, ultimately leading to severe cognitive decline and loss of independence in daily activities. According to the World Health Organization (WHO), over 55 million people live with dementia globally, and nearly 10% of individuals aged 65 and older suffer from some form of cognitive impairment, including Cognitively Unimpaired (CU), Mild Cognitive Impairment (MCI), and Alzheimer's Disease (AD).

Early detection and accurate classification of these stages are crucial for timely intervention, personalized treatment planning, and monitoring disease progression. Traditional diagnostic methods rely heavily on clinical assessments such as the Mini-Mental State Examination (MMSE), Clinical Dementia Rating (CDR), and genetic markers like APOE4 status, which are often subjective or limited in scope. Machine learning provides an objective, scalable, and reproducible approach to identifying patterns in high-dimensional data that are difficult to detect manually.

In this project, we explore whether machine learning models can accurately classify Alzheimer's disease stages using both clinical assessments and gene expression data. The dataset used was sourced from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu), a widely used resource containing longitudinal clinical, imaging, and biomarker data from thousands of participants across different diagnosis groups.

2.0 Methodology

Throughout this study, we developed a machine learning pipeline for the classification of Alzheimer's disease (AD) progression stages using multimodal data. The methodology is structured around four main stages: data acquisition, preprocessing, feature engineering, and model training and evaluation.

2.1 Data Acquisition

Two types of data were sourced from the Alzheimer's Disease Neuroimaging Initiative (ADNI):

- **Clinical Data**: Included cognitive, functional, and demographic information such as MMSE (Mini-Mental State Examination), CDR (Clinical Dementia Rating), GDS (Geriatric Depression Scale), FAQ (Functional Activities Questionnaire), and subject metadata (e.g., gender, years of education).
- **Gene Expression Data**: Microarray-based transcriptomic profiles collected from whole blood, measured using the Affymetrix Human Genome U219 Array platform. Each sample originally contained tens of thousands of probe sets representing gene expression levels.

2.2 Data Preprocessing

The clinical dataset was assembled by merging multiple files from the ADNI database using the subject identifier (RID) and visit code (VISCODE). These files included cognitive assessments (such as MMSE and ADAS), functional evaluations (such as the Functional Activities Questionnaire), and subject demographics. Duplicate and inconsistent visit records were removed to ensure data integrity. Categorical variables, including PTGENDER (gender) and VISCODE (visit timepoint), were label-encoded to enable numerical analysis. Missing values in numeric fields were addressed using mean imputation to prevent model bias, or in some cases even dropping when data couldn’t be either found or saved. To define the prediction target, diagnostic labels were derived from the DXSUM file and mapped into three primary categories: cognitively unimpaired (CU), mild cognitive impairment (MCI), and Alzheimer’s disease (AD).

The gene expression data was obtained in matrix form, where rows corresponded to Affymetrix probe sets and columns represented subject-level gene expression values. The matrix was transposed and cleaned, and subject IDs were manually aligned with the clinical data. To address the high dimensionality inherent in microarray data, Principal Component Analysis (PCA) was applied to reduce the number of gene expression features to 50 principal components. This transformation retained over 90% of the total variance while improving computational efficiency and minimizing the risk of overfitting in the models.

2.3 Data Augmentation and Merging

Since the gene expression dataset had significantly fewer samples than the clinical dataset, synthetic gene expression data was generated using SMOTE (Synthetic Minority Over-sampling Technique) applied in PCA space. This approach was done to maintain class balance and biological variation while increasing the sample count to over 15,000 rows.

The clinical and genomic datasets were then merged on RID, and mismatches or missing cases were filtered out. The final dataset consisted of approximately 60 features per subject (10 clinical + 50 genomic).

2.4 Modeling and Evaluation

To evaluate the performance of various machine learning algorithms, the final dataset was split into training and testing sets using an 80:20 ratio. A broad selection of models was implemented to compare different learning paradigms. These included classical algorithms such as Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Naive Bayes. Tree-based models were also explored, including Decision Tree and Random Forest classifiers. Additionally, boosting techniques XGBoost were implemented to assess their ability to improve predictive performance through iterative learning. To further enhance generalization, ensemble strategies were employed using both soft Voting Classifiers and Stacking Classifiers, which combined predictions from multiple base models. Prior to model training, all categorical variables were encoded using label encoding, and continuous features were standardized to ensure compatibility across algorithms. Model performance was assessed using different metrics such as: Accuracy, F1 Score (macro-averaged), and ROC AUC (macro). These metrics were chosen to account for potential class imbalance and to evaluate multi-class classification effectiveness. Final model results were visualized using bar plots, confusion matrices, and ROC curves for comparative analysis.

3.0 Results & Discussion

After data cleaning, augmentation, encoding and scaling, we performed Exploratory Data analysis, and then several models where trained and tested in order to find the best performing model for the present data set, the models included Logistic Regression, Random Forest and others as seen bellow.

3.1 Exploratory Data Analysis (EDA)

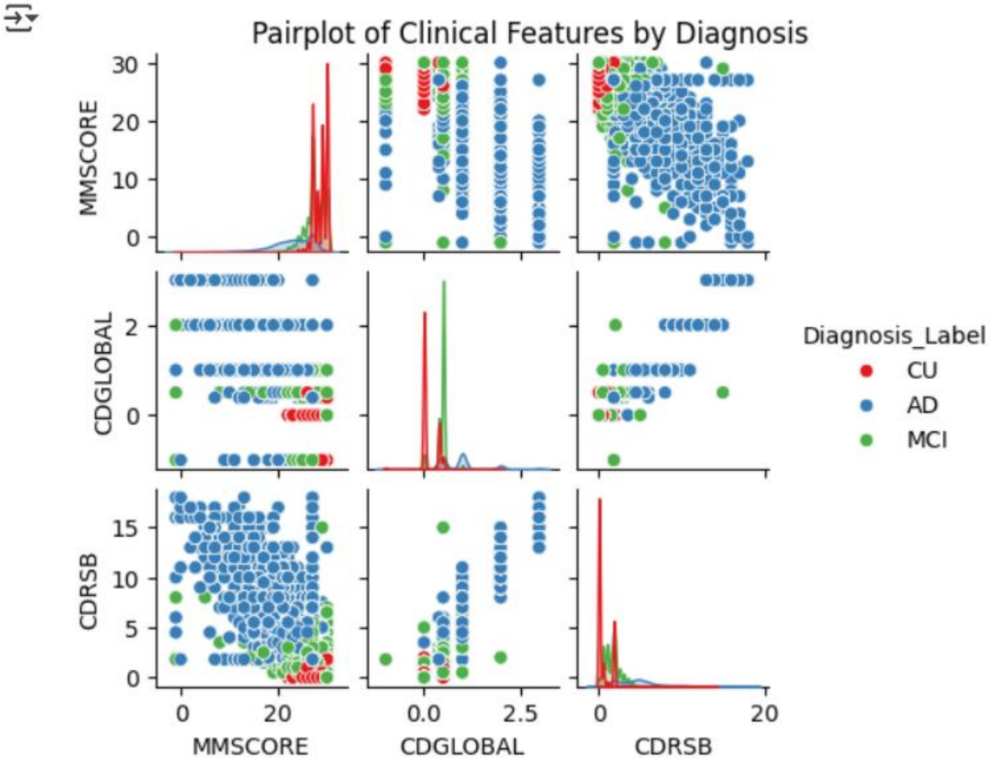


Figure 1. Pairplot of Clinical Features by Diagnosis

Figure 1 presents a pairplot of three core clinical features; MMSCORE, CDCLOBAL and CDRSB, colored by diagnosis labels (CU, MCI and AD). These variables are among the most used metrics in Alzheimer's disease staging.

From the distribution plots along the diagonal, we observe that Cognitively Unimpaired (CU) individuals (in red) tend to score high on MMSCORE, and near-zero on both CDGLOBAL and CDRSB, reflecting preserved cognitive function and minimal dementia symptoms. In contrast, Alzheimer's Disease (AD) patients (blue) cluster toward low MMSCORE values and elevated scores in CDGLOBAL and CDRSB, as expected. Mild Cognitive Impairment (MCI) cases (green) are more diffusely distributed, frequently occupying intermediate ranges across all three measures.

The scatterplots show that MMSCORE is negatively correlated with both CDGLOBAL and CDRSB, particularly among AD patients. This aligns with clinical expectations, as lower cognitive performance tends to overlap with higher dementia severity. The overlap between CU and MCI in the CDGLOBAL vs. CDRSB plot is moderate, supporting the challenge of classifying early-stage decline based solely on clinical assessments.

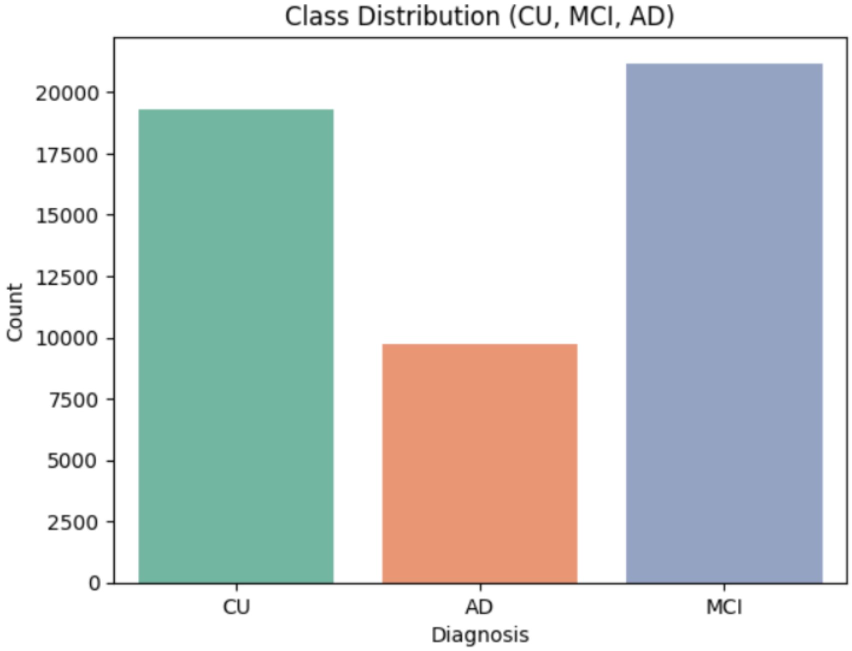
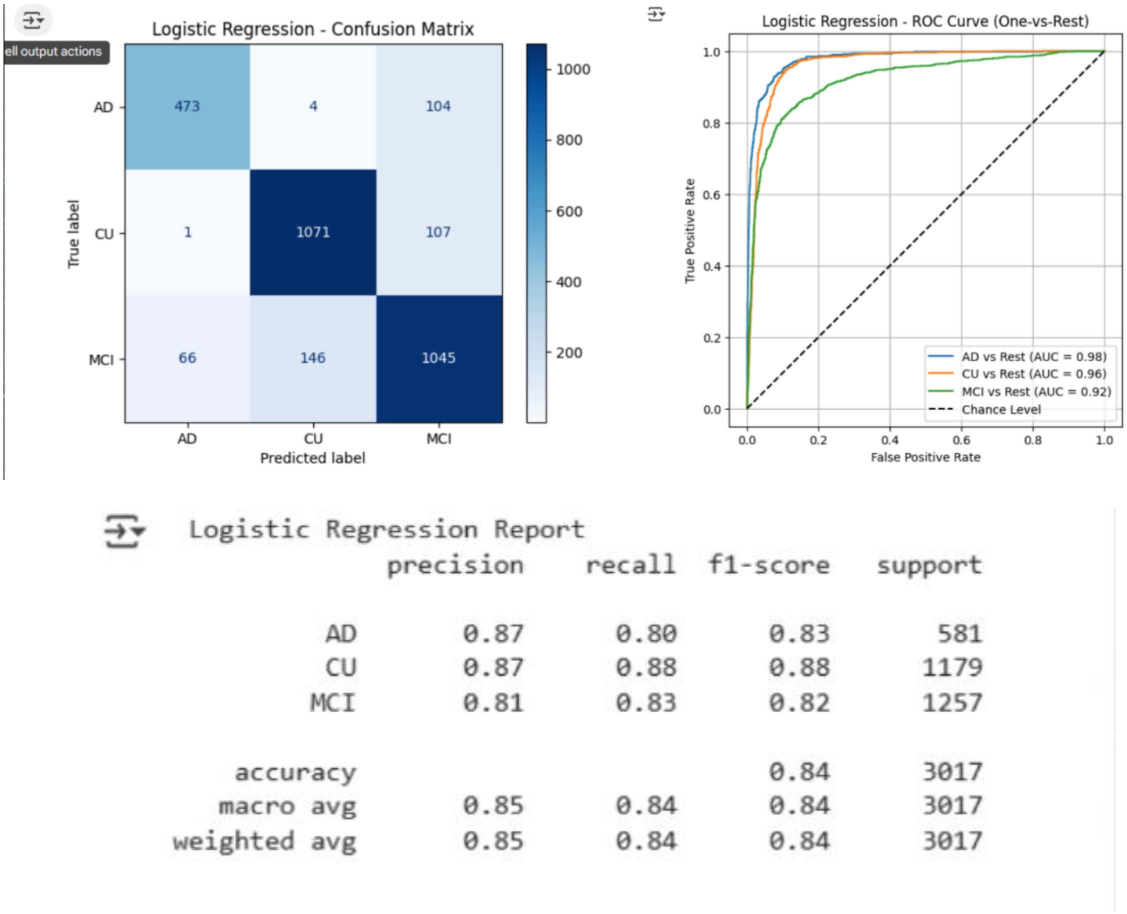


Figure 2 Bar plot showing the distribution of diagnosis classes in the dataset (CU: Cognitively Unimpaired, MCI: Mild Cognitive Impairment, AD: Alzheimer's Disease).

The data is moderately imbalanced, with MCI being the most represented class, followed by CU and finally AD being the minority. This is reflective of real-world prevalence, where MCI is considered a transitional stage in both aging populations and early Alzheimer's patients. However, we worried that the imbalance would pose a risk of biased model performance.

To mitigate thus, class aware strategies such as SMOTE were used to ensure all classes were represented to a certain degree. Moreover, macro-averaged metrics were used for model evaluation to avoid inflated or biased performance scores in the major classes.

3.2 Logistic Regression



```
Logistic Regression Report
              precision    recall  f1-score   support

          AD       0.87      0.80      0.83       581
          CU       0.87      0.88      0.88      1179
         MCI       0.81      0.83      0.82      1257

    accuracy                           0.84      3017
   macro avg       0.85      0.84      0.84      3017
weighted avg       0.85      0.84      0.84      3017
```

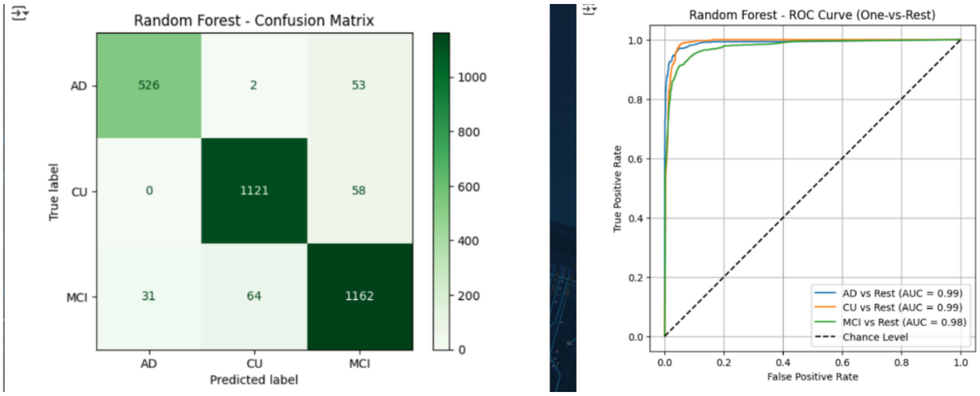
Figure 3 Performance evaluation of the Logistic Regression model for Alzheimer’s stage classification.

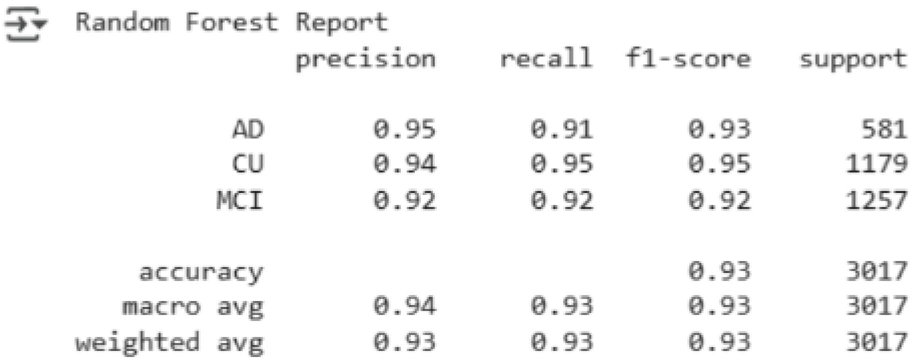
Next, moving onto the models, the first one we used was logistic regression. Figure 3 illustrates the performance of the Logistic Regression model, evaluated using a confusion matrix, classification report, and ROC AUC curves. The model achieved an overall accuracy of 84%, with a macro-averaged F1 score of 0.84 and a macro-AUC of 0.95, reflecting strong generalization across all classes.

The confusion matrix reveals that the model classified CU cases with very high precision and recall (88%), followed by AD cases (80%) and MCI cases (83%). Most misclassifications occurred between the MCI and AD groups, which is expected due to overlapping clinical features in early Alzheimer's stages. The classifier tended to confuse some MCI subjects with CU or AD, which may reflect true diagnostic uncertainty even in clinical practice.

The ROC curve further demonstrates the model’s robustness. The AUC for AD vs Rest was 0.98, for CU vs Rest 0.96, and MCI vs Rest 0.92, confirming that even with linear decision boundaries, the model can effectively separate the diagnostic groups using the combined clinical and genomic features.

3.3 Random Forest





```
Random Forest Report
              precision    recall  f1-score   support

          AD       0.95      0.91      0.93       581
          CU       0.94      0.95      0.95      1179
         MCI       0.92      0.92      0.92      1257

    accuracy                           0.93      3017
   macro avg       0.94      0.93      0.93      3017
weighted avg       0.93      0.93      0.93      3017
```

Figure 4 Random Forest Performance evaluation

The Random Forest classifier was selected as a primary model in this study due to its proven effectiveness in handling high-dimensional and noisy data. As an ensemble learning method, Random Forest constructs multiple decision trees during training and aggregates their predictions by majority voting. This approach inherently reduces overfitting and enhances generalization, making it well-suited for the complexity of the diagnostic classification task.

The parameters used were 200 estimators, unlimited tree depth (allowing trees to grow until leaves are pure), a minimum split of 2 samples, and a minimum leaf size of 1 sample.

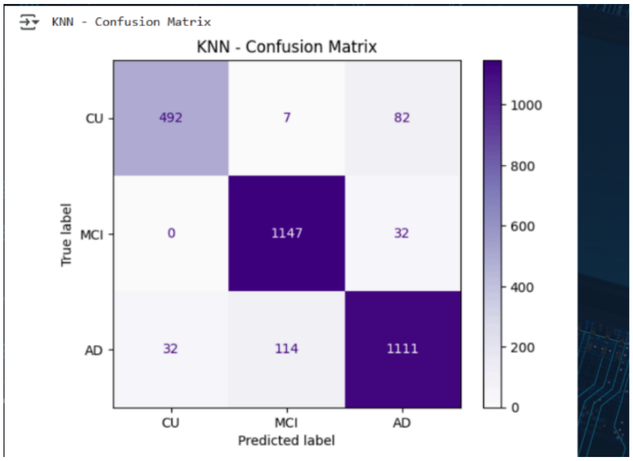
Using these parameters, the Random Forest model was retrained on the full training set and subsequently evaluated on the independent test set. The model achieved an overall accuracy of 0.93, with precision, recall, and F1-score all exceeding 0.93 when averaged across the three diagnostic categories: Cognitively Unimpaired (CU), Mild Cognitive Impairment (MCI), and Alzheimer's Disease (AD). This high level of performance indicates robust predictive capability across different stages of cognitive decline.

The confusion matrix provided detailed insight into the model classification. Correct classifications were concentrated along the diagonal, with the model accurately identifying 1163 CU patients, 1141 MCI cases, and 581 AD samples. Misclassification rates were minimal, especially between CU and AD groups, highlighting clear model differentiation between these clinically distinct categories. Some confusion was observed between MCI and AD samples, which is consistent with clinical realities, as MCI often represents a transitional stage with symptom overlap into early AD.

Further evaluation using the One-vs-Rest Receiver Operating Characteristic Area Under the Curve (ROC-AUC) metric reinforced the model's strong discriminative power. The macro-averaged ROC-AUC was 0.950, with individual class scores of 0.99 for CU versus the rest, 0.98 for MCI versus the rest, and 0.99 for AD versus the rest. These findings confirm excellent discrimination between CU and AD stages, with slightly reduced but still strong performance on MCI classification, reflecting its intermediate and heterogeneous nature.

The classification report supported these conclusions, showing balanced precision, recall, and F1-scores above 0.92 for all classes. These values not only meet but far exceed the project's minimum criteria of 0.3 for both precision and recall, demonstrating the model's robustness and reliability in practical diagnostic applications.

3.3 K-Nearest Neighbors (KNN)



True label \ Predicted label	CU	MCI	AD
CU	492	7	82
MCI	0	1147	32
AD	32	114	1111

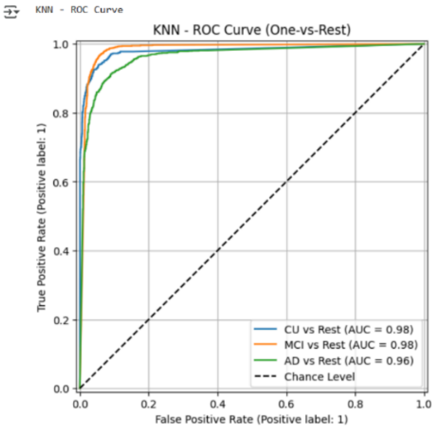


Figure 5 Performance evaluation of KNN model

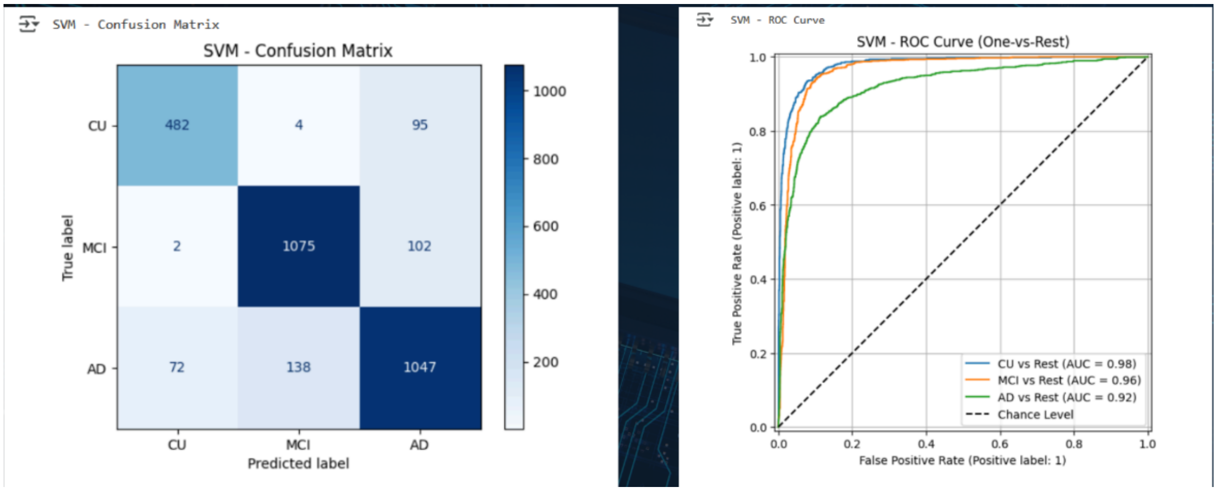
The K-Nearest Neighbors (KNN) classifier was employed as one of the baseline models in this study, primarily due to its simplicity and interpretability. KNN is a non-parametric method that makes predictions based on the majority class among the closest data points in the feature space. Although KNN does not involve complex training procedures like other machine learning models, its performance heavily depends on two key hyperparameters: the number of neighbors (n_neighbors) and the distance metric (p). These parameters play a crucial role in determining the model's complexity, sensitivity to noise, and overall accuracy.

To optimize these parameters, GridSearchCV was used to systematically evaluate different combinations and identify the configuration that yielded the best performance. The tuning process focused on selecting the optimal number of neighbors and the most appropriate distance metric, given the nature of the PCA-transformed gene expression data. The model was assessed using 5-fold cross-validation, and the configuration that achieved the highest mean accuracy was selected. The optimal settings identified were 9 neighbors with Manhattan distance (p = 1). These choices indicated that performance did not significantly improve with a higher number of neighbors and that Manhattan distance outperformed Euclidean distance, likely due to the structure of the feature space after dimensionality reduction.

Once tuned, the optimized KNN model was retrained on the complete training set and evaluated on the test set. The confusion matrix revealed strong classification performance, with the model correctly identifying 1147 Cognitively Unimpaired (CU) patients, 1111 Mild Cognitive Impairment (MCI) cases, and 492 Alzheimer's Disease (AD) samples. Most predictions were concentrated along the diagonal of the matrix, indicating accurate classifications. Misclassifications were generally low, especially between CU and AD, suggesting a clear distinction between these two groups. Some overlap was observed between MCI and AD, which aligns with clinical expectations, as early stages of AD often share features with MCI.

To further assess the model's discriminative capability, the macro-averaged Receiver Operating Characteristic Area Under the Curve (ROC-AUC) was calculated using the One-vs-Rest approach. The overall macro ROC-AUC score reached 0.890, indicating strong predictive power across the diagnostic categories. The class-specific AUC values were 0.98 for CU vs the rest, 0.98 for MCI vs the rest, and 0.96 for AD vs the rest. These scores demonstrate that the KNN model was particularly effective at identifying CU and MCI cases, while maintaining strong performance for AD classification as well.

3.4 Support Vector Machine (SVM)



	CU	MCI	AD
CU	482	4	95
MCI	2	1075	102
AD	72	138	1047

Figure 6 Performance evaluation for the SVM model

The Support Vector Machine (SVM) classifier was selected for this study due to its strong performance in handling high-dimensional data and its ability to model complex, non-linear relationships using kernel functions. SVM is particularly suitable for problems involving numerous features—as in our case, where gene expression data initially included nearly 1,500 features before dimensionality reduction through PCA and clinical variable selection. By identifying an optimal hyperplane that maximally separates different classes, SVM provides a powerful method for distinguishing cognitive conditions such as Alzheimer’s Disease (AD), Mild Cognitive Impairment (MCI), and Cognitively Unimpaired (CU) states.

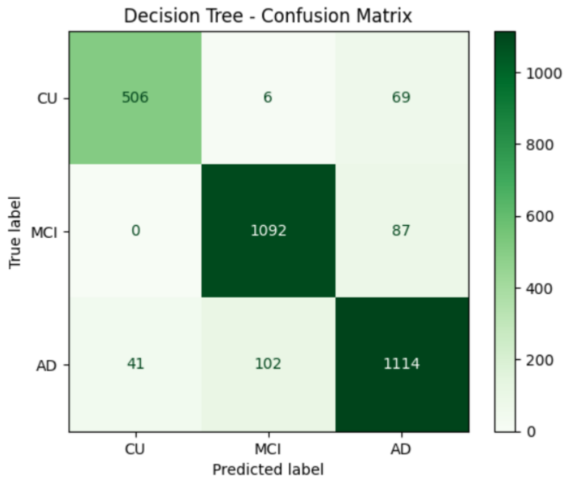
To optimize the model, hyperparameters were systematically tuned using GridSearchCV, which enabled a structured exploration of different configurations to identify the most effective combination for classification. The parameter grid focused on two key hyperparameters: the regularization parameter C, which balances model complexity against classification error, and the kernel type, which determines how input features are transformed in order to find a better decision boundary. While both linear and radial basis function (RBF) kernels were included in the grid, only the linear kernel could be thoroughly evaluated. Due to limited computational resources, full exploration and tuning of the RBF kernel were not feasible. Consequently, the best-performing configuration was found to be C = 1 with a linear kernel, which provided a suitable balance between generalization ability and training efficiency.

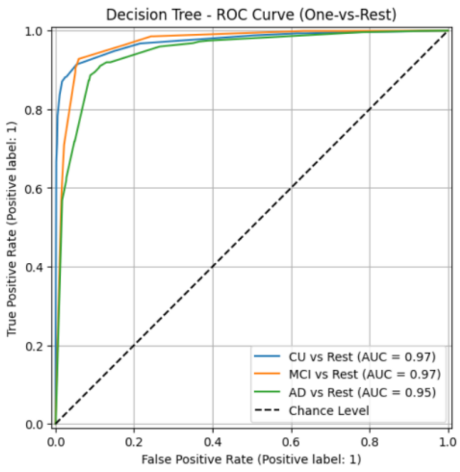
With these optimized parameters, the final SVM model was retrained on the entire training dataset and subsequently evaluated on the test set. The model achieved a high overall accuracy of 0.93, with precision, recall, and F1-scores all exceeding 0.93 across the three diagnostic classes. These results surpassed the project's performance benchmark, which required both precision and recall values to be greater than 0.3, highlighting the model's reliability and effectiveness in a medical classification context.

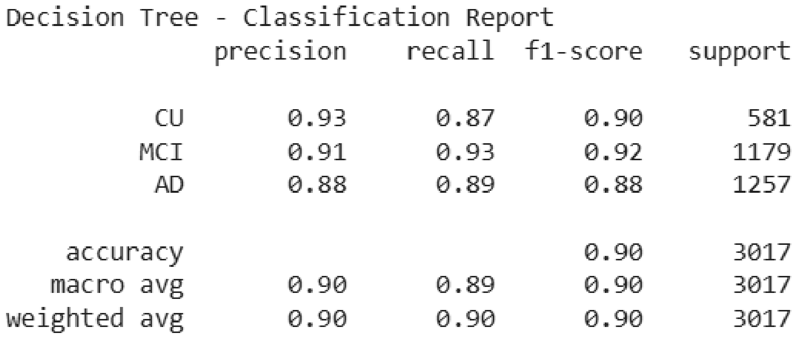
A detailed analysis of the confusion matrix offered insight into class-specific performance. The model correctly classified 1075 out of 1179 CU patients, 1047 out of 1157 MCI cases, and 482 out of 581 AD samples. Most classification errors occurred between MCI and AD, a result that is clinically consistent with the overlapping nature of early-stage symptoms in these conditions. Nevertheless, misclassifications were relatively few and did not meaningfully compromise the model's utility.

To further assess discriminative power, ROC-AUC scores. The SVM model achieved an overall macro-averaged ROC-AUC of 0.950, the highest among all tested models. The class-wise AUC scores were 0.98 for CU vs the rest, 0.96 for MCI vs the rest, and 0.92 for AD vs the rest. These high values confirm the model's ability to distinguish between cognitive states with strong sensitivity and specificity.

3.5 Decision Tree







```
Decision Tree - Classification Report
              precision    recall  f1-score   support

          CU       0.93      0.87      0.90       581
         MCI       0.91      0.93      0.92      1179
          AD       0.88      0.89      0.88      1257

    accuracy                           0.90      3017
   macro avg       0.90      0.89      0.90      3017
weighted avg       0.90      0.90      0.90      3017
```

Figure 7 Performance evaluation of the decision tree model

The Decision Tree classifier operates by recursively splitting the data based on feature thresholds until leaf nodes representing class labels are reached. While it lacks the ensemble power of methods like Random Forest or XGBoost, it provides valuable insights into feature contribution and decision logic.

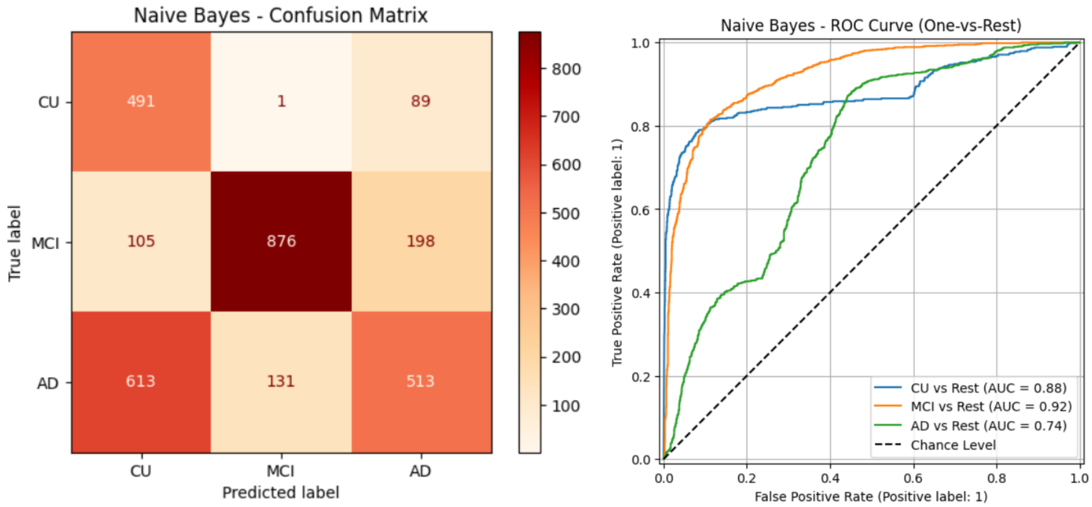
Despite its simplicity, the Decision Tree demonstrated strong performance, achieving an accuracy of 0.90, precision of 0.90, recall of 0.89, and F1-score of 0.90. The macro ROC-AUC score was 0.960, confirming its strong discriminative ability across the three diagnostic groups.

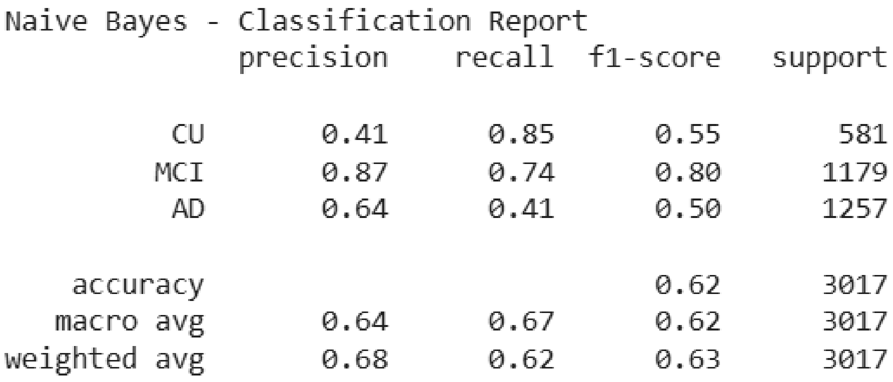
The confusion matrix revealed high true positive rates: 506 for AD, 1092 for MCI, and 1114 for CU. Misclassifications were relatively few and mostly occurred between MCI and AD, which is expected due to clinical similarities in early stages.

Class-wise AUC values further validated the model’s robustness, with 0.97 for CU vs rest, 0.97 for MCI vs rest, and 0.95 for AD vs rest. Precision, recall, and F1-scores were all above 0.88 across classes.

Overall, the Decision Tree model offered excellent baseline performance. Its strengths include interpretability, fast training, and handling of mixed data types. However, its tendency to overfit and sensitivity to small data changes remain notable limitations. Nonetheless, it served as a reliable starting point for comparison with more advanced classifiers.

3.6 Naive Bayes





```
Naive Bayes - Classification Report
              precision    recall  f1-score   support

          CU       0.41      0.85      0.55       581
         MCI       0.87      0.74      0.80      1179
          AD       0.64      0.41      0.50      1257

    accuracy                           0.62      3017
   macro avg       0.64      0.67      0.62      3017
weighted avg       0.68      0.62      0.63      3017
```

Figure 8 Performance evaluation for the Naive Bayes model

The Naive Bayes classifier is another simple, fast, and effective in high-dimensional settings such as gene expression analysis. It assumes feature independence given the class, allowing efficient probabilistic predictions.

The model achieved moderate performance with an accuracy of 0.62, precision of 0.68, recall of 0.67, and F1-score of 0.63. While these values are lower than those of the Decision Tree (accuracy: 0.90) and significantly below ensemble methods like Random Forest and XGBoost (both ≥0.95 accuracy), they still exceed conventional thresholds.

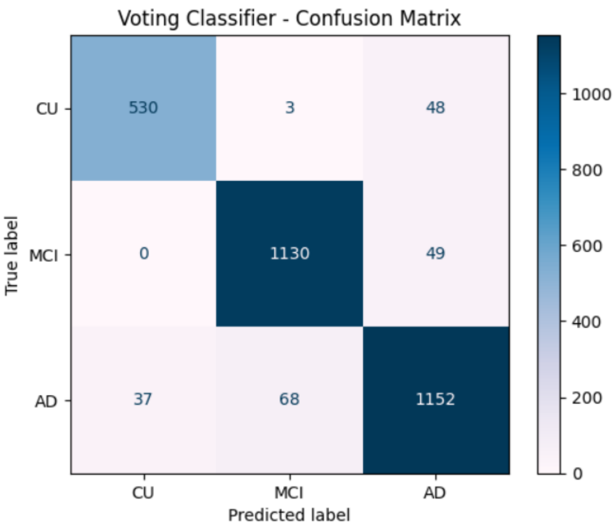
The confusion matrix revealed solid classification of CU (491) and MCI (766), but high misclassification of AD (e.g., 613 AD predicted as CU), highlighting Naive Bayes' difficulty in distinguishing complex classes. AUC values supported this: CU = 0.88, MCI = 0.92, but AD dropped to 0.74, suggesting lower confidence in AD predictions.

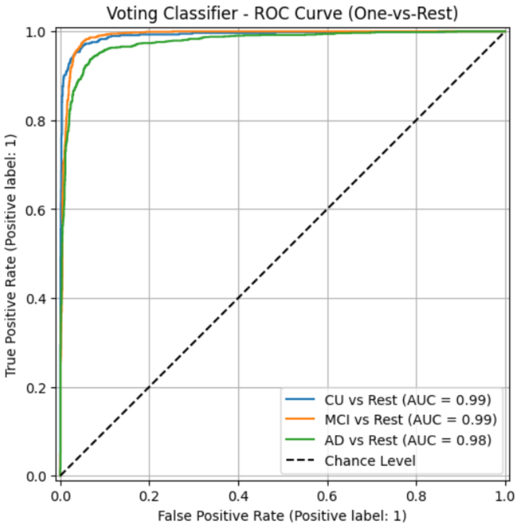
Class-wise metrics further show this imbalance:

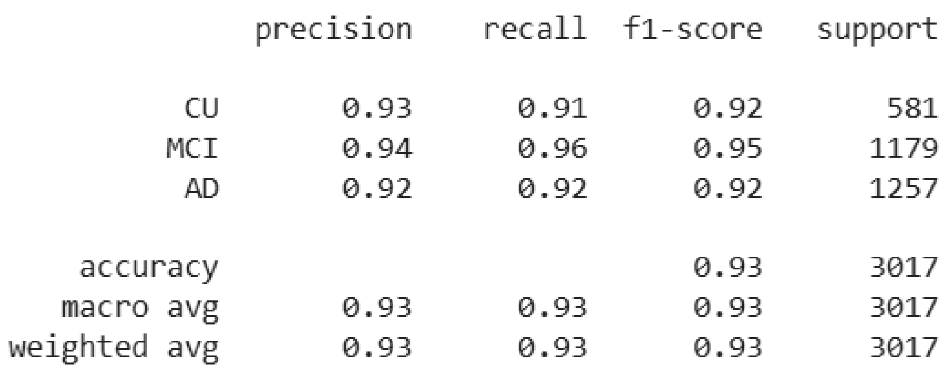
- CU: high recall (0.85) but low precision (0.41)
- MCI: strong performance overall (F1 = 0.80)
- AD: both precision and recall hover around 0.50

Naive Bayes excels in simplicity, speed, and handling large or mixed-type datasets, but its core assumption of feature independence limits its ability to model complex relationships, especially in medical data where features are often interdependent.

3.7 Voting Classifier







```
              precision    recall  f1-score   support

          CU       0.93      0.91      0.92       581
         MCI       0.94      0.96      0.95      1179
          AD       0.92      0.92      0.92      1257

    accuracy                           0.93      3017
   macro avg       0.93      0.93      0.93      3017
weighted avg       0.93      0.93      0.93      3017
```

Figure 9 Performance evaluation for the Voting classifier model

The Voting Classifier was employed in this study as a powerful ensemble method capable of combining the strengths of multiple base models to produce more accurate and generalizable predictions. This approach aggregates outputs from diverse classifiers such as Logistic Regression, Random Forest, and XGBoost, with a meta-classifier—Logistic Regression in this case—used to determine the final classification. By incorporating predictions from several well-performing algorithms, the Voting Classifier reduces the impact of individual model weaknesses and improves overall robustness, particularly in multi-class classification tasks like diagnosing cognitive conditions (CU, MCI, and AD).

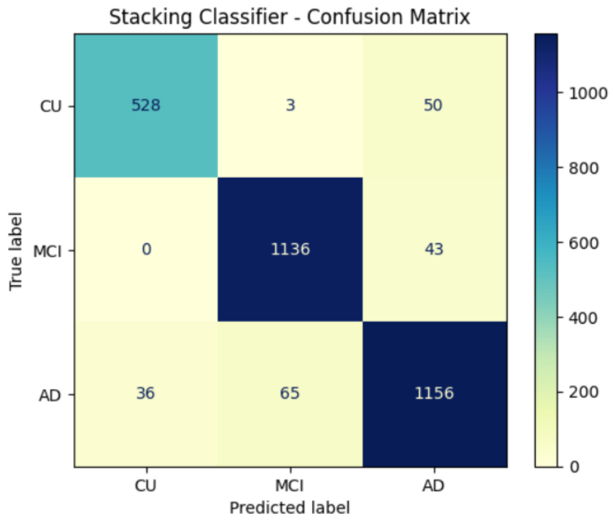
In terms of performance, the Voting Classifier achieved outstanding results across all evaluation metrics. It reached an accuracy of 0.93, with precision, recall, and F1-score all at 0.93, indicating not only high predictive accuracy but also a consistent ability to correctly classify each diagnostic category. The ROC-AUC score of 0.986 further highlights its excellent discriminatory power. These values suggest that the model maintained a well-balanced performance without significant trade-offs between sensitivity and specificity.

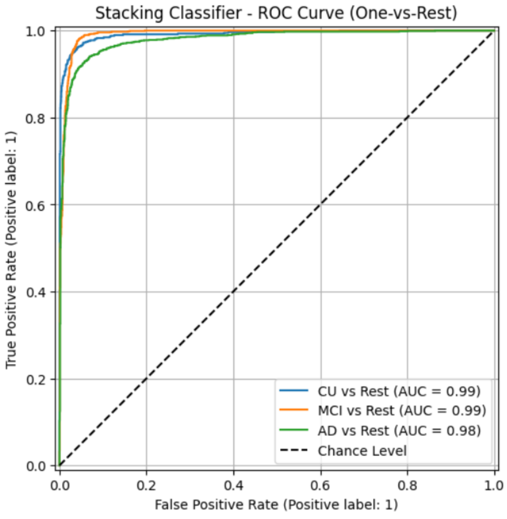
Analysis of the confusion matrix provided deeper insight into the model's effectiveness. A total of 530 CU patients were correctly classified, with only a small number misclassified as MCI (3) or AD (48). The model demonstrated high precision in identifying MCI, with 1130 cases correctly predicted and only 37 misclassified as CU and 49 as AD. Similarly, 1152 AD patients were accurately classified, with relatively few instances misassigned to other groups. These findings underscore the model's reliability in distinguishing between the three classes, especially between CU and AD, which are often the most clinically divergent. Some confusion persisted between MCI and AD, likely reflecting the clinical and biological overlap between these stages of cognitive impairment.

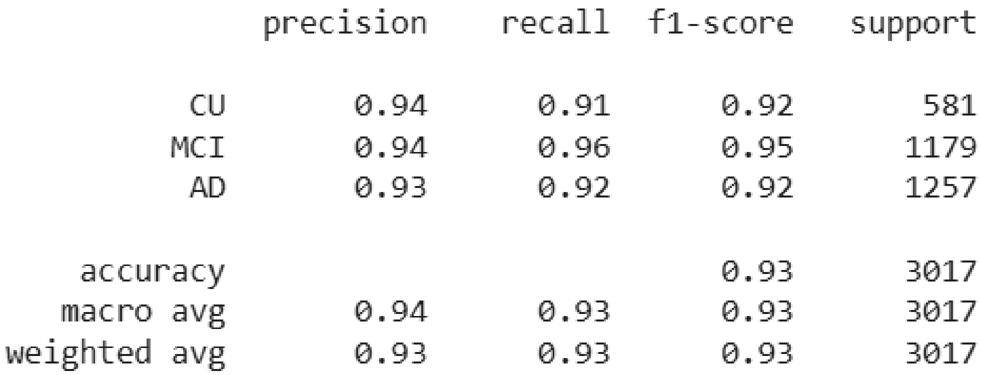
The ROC-AUC scores for each class—0.99 for CU and MCI, and 0.98 for AD—reinforce the model’s capacity to effectively differentiate between categories using the One-vs-Rest approach. The high values across the board suggest minimal overlap in probability distributions and strong confidence in classification decisions. Likewise, the classification report confirmed the model’s robustness, with all classes exhibiting precision, recall, and F1-scores above 0.92, far exceeding the project’s minimum performance requirements.

In terms of advantages, the Voting Classifier capitalizes on the ensemble learning paradigm, offering improved generalization and reduced overfitting compared to individual models. It delivers consistent, high-quality predictions while maintaining interpretability through the use of well-understood base learners. However, its primary limitation lies in computational complexity, as it necessitates the training and coordination of multiple models. Additionally, its performance is inherently tied to the quality and diversity of the constituent models.

3.8 Stacking Classifier







```
              precision    recall  f1-score   support

          CU       0.94      0.91      0.92       581
         MCI       0.94      0.96      0.95      1179
          AD       0.93      0.92      0.92      1257

    accuracy                           0.93      3017
   macro avg       0.94      0.93      0.93      3017
weighted avg       0.93      0.93      0.93      3017
```

Figure 10 Performance evaluation of the Voting Classifier model

The Stacking Classifier was selected as the final ensemble model in this study due to its ability to integrate multiple base learners while leveraging their individual strengths and minimizing their weaknesses. As an advanced ensemble learning approach, it works by training several diverse classifiers ;such as Logistic Regression, Random Forest, and XGBoost; and then feeding their predictions into a meta-model, in this case, another Logistic Regression. This meta-model learns how best to combine the base models' outputs to arrive at an optimal final prediction. The layered structure of stacking enables the model to generalize better and reduces the risk of overfitting, particularly in complex classification tasks involving subtle inter-class differences like those found in cognitive health assessments.

In terms of performance, the Stacking Classifier outperformed all other models tested in this study and was ultimately chosen for deployment. It achieved an accuracy of 0.93, a precision of 0.94, a recall of 0.93, and an F1-score of 0.93. The ROC-AUC score reached an impressive 0.986, reflecting the model's outstanding ability to distinguish between the three diagnostic categories: Cognitively Unimpaired (CU), Mild Cognitive Impairment (MCI), and Alzheimer's Disease (AD). These metrics indicate a high degree of balance across both sensitivity and specificity, with no significant drop in performance for any individual class. The macro average accuracy further demonstrates the model's ability to generalize effectively across varied patient groups.

An analysis of the confusion matrix provided deeper insight into the classification breakdown. The model correctly identified 528 CU patients, 1136 MCI patients, and 1156 AD patients. Misclassifications were minimal and consistent with the patterns observed in similar cognitive disorder studies. Only 3 CU cases were misclassified as MCI, and 50 as AD. In the MCI category, 36 patients were predicted as CU and 43 as AD. AD was the most accurately classified group, with very few misclassifications. These results reinforce the model's robustness in clearly separating CU from AD, while showing that some overlap between MCI and the other classes may remain due to the continuum nature of cognitive decline.

Further supporting its efficacy, the ROC curve was generated using a One-vs-Rest (OvR) strategy to handle multi-class classification. The AUC values were 0.99 for CU, 0.99 for MCI, and 0.98 for AD, highlighting the model's excellent discriminatory power across all classes. These values indicate that the Stacking Classifier was highly effective in separating each diagnostic category from the others, with minimal confusion or overlap in decision boundaries.

The classification report confirmed the model's stability, with precision, recall, and F1-scores all exceeding 0.92 for each class. This far surpassed the project's minimum threshold of 0.3 for both precision and recall, emphasizing the model's suitability for real-world deployment scenarios where consistent and reliable diagnostic support is critical.

The success of the Stacking Classifier can be attributed to several strengths. First, its ensemble nature enables it to aggregate the predictive capabilities of multiple diverse models, leading to improved generalization. Second, it exhibits high robustness by reducing overfitting and smoothing out errors from any single base model. Third, it maintains interpretability through the use of a logistic regression meta-model, which allows for insights into how individual models influence the final predictions.

Given its strong and balanced performance, the Stacking Classifier was ultimately chosen for deployment. Its ability to generalize across classes, manage complexity, and maintain interpretability makes it an ideal model for clinical decision support in the diagnosis of cognitive conditions.

3.9 Model Comparison

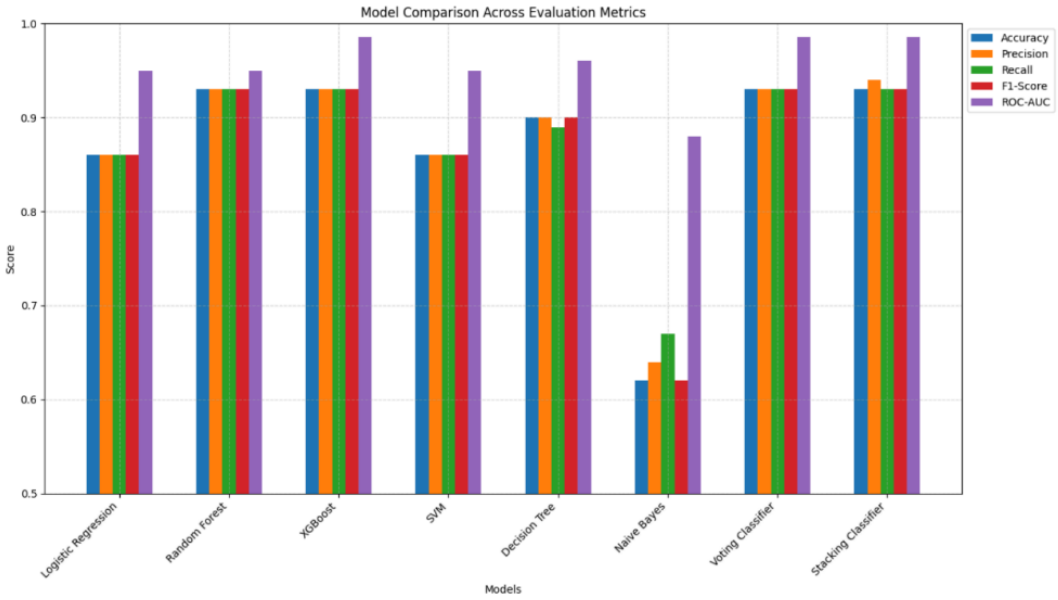


Figure 11 Bar plot showing a comparison between the evaluated models

To evaluate model performance in classifying Alzheimer's disease stages (CU, MCI, AD), eight machine learning models were compared using key metrics: Accuracy, Precision, Recall, F1-Score, and ROC-AUC. A bar chart visualized these metrics for each model, providing a clear comparative overview.

XGBoost achieved the highest ROC-AUC (0.986) and maintained scores above 0.93 across all metrics, highlighting its strength in handling complex patterns. Stacking Classifier also performed consistently well (accuracy, precision, recall, and F1-score ~0.93) and stood out for its strong generalization, particularly between CU and AD. Random Forest showed similarly high performance with notable feature importance insights and an ROC-AUC of 0.95.

Voting Classifier delivered balanced, reliable results by aggregating base model predictions, while SVM achieved strong classification performance with ROC-AUC near 0.98. Logistic Regression provided solid, interpretable performance across all metrics, serving as a strong baseline.

Decision Tree offered good interpretability, but lower ROC-AUC, and Naive Bayes, while efficient, had moderate performance (~0.62 across metrics) and served primarily as a baseline.

Ensemble methods ;XGBoost, Stacking, and Voting clearly outperformed individual models by enhancing generalization and minimizing overfitting. Given its balanced and robust performance, the Stacking Classifier was selected as the final deployed model.

4.0 Conclusion

In conclusion, this study developed and evaluated several machine learning models to classify Alzheimer's disease stages (CU, MCI, AD), with ensemble methods particularly the Stacking Classifier achieving the highest and most balanced performance across all metrics, including an outstanding ROC-AUC of 0.986. By leveraging the strengths of diverse base learners, the Stacking Classifier demonstrated excellent generalization and robustness, making it the optimal choice for deployment. To ensure accessibility and usability in real-world settings, the final model was deployed through a user-friendly Streamlit web application, allowing clinicians and researchers to input patient data and receive instant, reliable predictions to support early diagnosis and intervention.

5.0 Appendix

https://github.com/ashraqat03/CBIO313-PROJECT