Chapter 8

Leveraging handwriting dynamics, explainable AI and Machine learning for Alzheimer prediction

In previous chapters, we explored dynamic hand gestures, specifically those related to handwriting, and discovered that these dynamics could help predict neurological diseases like Alzheimer's. By employing explainable AI, we aim to deepen our understanding of the features and dynamics involved. In this study, a benchmark dataset, "DARWIN," is selected to investigate the correlation between motion patterns and disease detection. The objective is to leverage this knowledge to enhance our future sEMG-based research. By doing so, we aim to build a robust framework that facilitates early diagnosis as well as developing a comprehensive understanding of Alzheimer's disease through the analysis of fine motor skills.

Section 8.1 provides an introduction including major contributions. The experimental setup is given in Section 8.2, highlighting the Materials and methods used, including the dataset, data preprocessing, and proposed stacked ensemble method. In Section 8.3, we showcase our results, offering an in-depth analysis of performance

metrics and incorporating explainable AI. Lastly, Section 8.4 consolidates our primary conclusions and outlines avenues for future improvements.

8.1 Introduction

Alzheimer's Disease (AD), a neurological condition, has captured the global health community's attention due to its adverse effect on daily human life [301]. Alzheimer's is progressive and irreversible in nature, making it one of the most feared neurodegenerative conditions as it erodes the core facets of human cognition: memory, thinking, and behavior [255]. Moreover, there has been a staggering growth from 20.3 million affected individuals in 1990 to a predicted 43.8 million by 2050 [168]. This fact underscores the imminent health crisis it projects.

As the understanding of AD has grown, much emphasis has been provided on early diagnosis. Capturing the disease in its early stages, potentially even before the emergence of overt dementia symptoms, has shown improved patient outcomes [302]. However, achieving early diagnosis requires effective methodologies to decipher the intricate neural patterns and subtle manifestations associated with early-stage AD. [6].

Machine learning has revolutionized many sectors, including health diagnostics [303]. With their layered architectures and ability to identify complex relationships within data, machine learning models have shown significant promise [304]. Utilizing handwriting dynamics, alongside machine learning, is a pivotal tool in predicting Alzheimer's disease. By analyzing intricate patterns in handwriting, machine learning algorithms can detect subtle irregularities and changes over time that may signal the onset of AD.

The integration of explainable AI approaches enhances the utility of machine learning models by providing clear insights into their decision-making processes [305]. Explainable AI fosters trust and understanding among clinicians and patients by

revealing the rationale behind model predictions. This is particularly important in healthcare, where transparency is crucial for the adoption of new technologies.

In this study, we used explainable AI principles, specifically SHAP (SHapley Additive exPlanations), to provide insights into the features responsible for AD [184]. SHAP helps identify the most influential features and explains the decision-making process, enhancing trust and actionable insights for clinical use. Additionally, we applied the stacking ensemble approach to explore its potential for AD classification, combining multiple predictive models to improve accuracy and robustness.

Ensemble stacking is a machine learning technique that combines multiple predictive models to improve accuracy [306]. By layering diverse models, stacking leverages their strengths and compensates for their weaknesses, creating a more robust and dependable classification system. Given the complex and varied nature of AD, with its many symptoms and overlapping conditions, a stacking ensemble holds significant potential. It not only aids in building effective diagnostic tools but also enhances the accuracy essential for early-stage detection.

8.1.1 Major contribution of the work

The primary contributions of this chapter can be summarized as follows:

1. We employed the advanced explainable AI technique, SHAP (SHapley Additive exPlanations), to provide clear insights into the features responsible for Alzheimer's disease. This approach enhances model transparency and interpretability, allowing us to pinpoint pivotal features and understand their contributions to the classification process.

2. We introduced an effective stacking ensemble model tailored to classify Alzheimer's disease based on handwriting movements. This model combines multiple predictive algorithms to improve precision and robustness.

3. Our model was tested against a benchmark dataset and demonstrated performance metrics comparable to state-of-the-art techniques.

8.2 Methods and Materials

8.2.1 Dataset

We used the DARWIN [168], a benchmark dataset that collects different information related to handwriting dynamics. The details of the dataset are provided in Section 2.3.10.

8.2.2 Proposed Model: Stacking ensemble Classifier for Alzheimer's Disease Detection

With the motive of enhancing the accuracy of early detection of Alzheimer's Disease (AD), we adopted the efficient "stacking" technique. Our model is built upon a rich foundation of diverse machine-learning algorithms. This includes CatBoost, a gradient boosting library [307]; Random Forest [308], which integrates numerous decision trees for its predictions; Naive Bayes that calculates the likelihood of an event based on Bayes theorem [309] ; K-Nearest Neighbors (KNN) that classifies data points based on the majority class of its nearest neighbors [310]; and Linear Discriminant Analysis (LDA), focused on determining attributes that capture the most variance between classes [311]. These base models were then fed their predictions into another machine learning model designated as the meta-model, in our case, another RandomForest. By processing the combined insights of the base models, this meta-model offers the final prediction. Essentially, the predictions from the base models are "stacked", serving as input for our meta-model, which then deduces the final output. The efficacy of our approach is gauged by testing its predictions against actual outcomes, and its inherent strength lies in leveraging the capabilities of multiple models, aiming for superior prediction accuracy and reliability. Figure 8.1 outlines the proposed ensemble architecture.

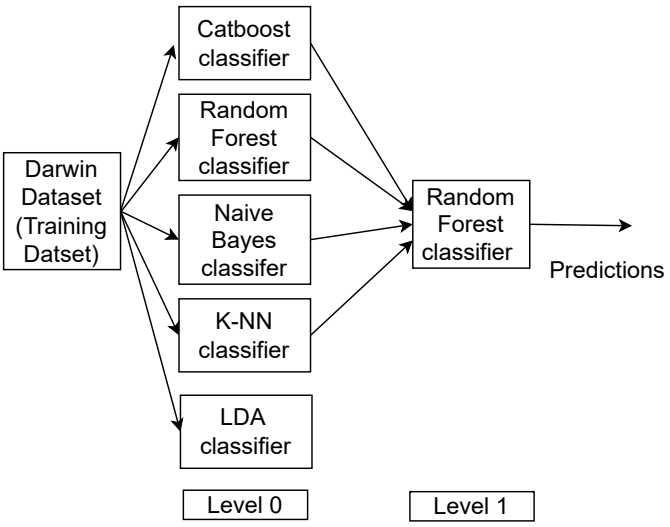

FIGURE 8.1: Schematic diagram showing proposed stack ensemble model

8.2.3 Shapley Additive Explanations for Alzheimer's Disease Detection: An explainable AI approach

Interpreting the results and understanding the decision-making process of machine learning models has become paramount, especially in critical fields like medical diagnostics. In the context of Alzheimer's Disease Detection, understanding which features or biological markers most influence a model's prediction can provide invaluable insights for medical professionals.

SHAP (Shapley Additive Explanations) is one such technique that offers a window into the model's internal workings [184]. Drawing its foundations from game theory's Shapley values, SHAP values distribute the 'credit' for a prediction among the input features [185]. In essence, they quantify how each feature contributes to a particular prediction compared to a baseline prediction.

In the realm of Alzheimer's Disease Detection, ranking features based on their SHAP values can spotlight the most critical biomarkers or indicators influencing the model's predictions. Visualization tools, like SHAP summary plots, further transform these values into intuitive graphics, elucidating feature importance and

aiding healthcare practitioners in understanding the predictions' underlying rationale [312, 313]. Fig. 8.3 highlights the features with maximum impact on the class prediction for Alzheimer's classification using the Darwin dataset.

8.3 Results & Discussion

8.3.1 Performance Measures

Our primary objective in conducting these experiments was to gauge the potency of our proposed two-tiered stacking ensemble technique against the widely recognized DARWIN dataset.

Performance metrics are fundamental indicators of a model's efficacy. In our evaluation, we focus on standard metrics such as Accuracy, MCC, F1 score, Precision, and Recall, as these are crucial for medical diagnoses like Alzheimer's. Impressively, our ensemble model reported a precision of 94.44% in Alzheimer's detection. Complementary to this, the model also recorded commendable scores of 88.57% in Accuracy, 77.56% in MCC, and 85% in Recall when tested against the DARWIN dataset. When compared with existing state-of-the-art methodologies on the same dataset, our model's metrics are comparable but, in some aspects, superior. An integral metric that further underscores our model's performance is the AUC-ROC score, which stands at 0.92, indicating a highly reliable model in distinguishing between the positive and negative classes. Figure 8.2 highlights the ROC AUC curve for the DARWIN dataset, while Table 8.1 presents a summary of the performance metrics achieved for the DARWIN dataset.

TABLE 8.1: Performance metrics achieved for our proposed stacking ensemble model

Stacking Accuracy	Stacking MCC	Stacking Precision	Stacking Recall	Stacking F1 Score
0.8857	0.7756	0.9444	0.8500	0.8947

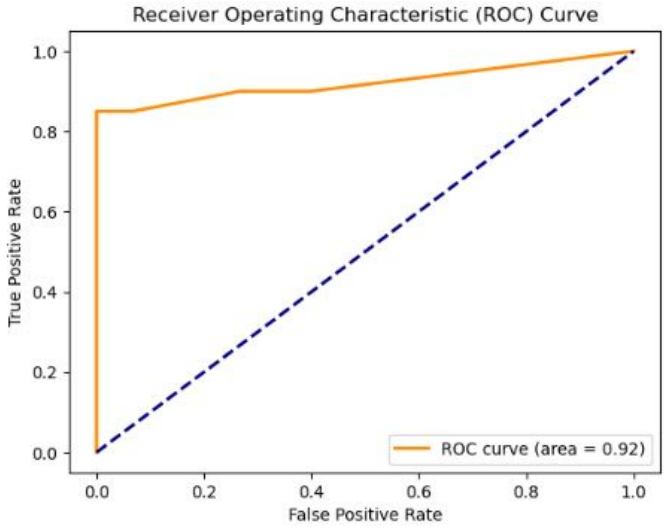

FIGURE 8.2: ROC AUC curve generated using machine learning ensemble

8.3.2 Feature interpretability and ranking using SHAP

The SHAP (SHapley Additive exPlanations) analysis provides valuable insights into the impact of various features on the model's output [184]. By visualizing these impacts, we can better understand which features are most influential in predicting Alzheimer's disease (AD) based on handwriting movements. The following sections discuss the key observations from the SHAP plots and their implications for our study. Figure 8.3 displays a bar chart that represents the average magnitude of SHAP values for each handwriting feature across all instances in the dataset. This plot provides a high-level overview of the importance of each feature in making a prediction. Features are listed on the y-axis of the plot, while the x-axis represents the average magnitude of the SHAP values for each feature. The magnitude signifies the average impact of each feature on the model's output, regardless of the direction (positive or negative) of that impact. For our work, it signifies the presence of AD.

Features are sorted by importance, with the most impactful feature at the top. Importance is determined by the average absolute SHAP value for each feature. Thus, the feature at the top has the most significant impact on the model's predictions. Considering Figure 8.3, the count of pen-down performed for task 19 (i.e., while copying the postal order fields) is the most influential feature globally that

helps our ensemble model predict the presence of AD in the subject. A possible explanation is that frequent pen-down events may reflect microtremors or interruptions in writing, which can be indicative of motor control issues associated with AD.

The next influential feature is the total time duration required for task 23 (i.e., Write the dictated phone number). Longer task completion times (higher feature values) are associated with higher SHAP values, suggesting an increased likelihood of an AD prediction. This aligns with the understanding that individuals with AD may experience cognitive slowing or difficulty in motor planning. The SHAP value plot shows a clear positive correlation between total time and SHAP values, indicating its significant role in the model.

Air time for task 5 (i.e., Retrace 3 cm diameter circle four times) and task 17 (i.e. Copy six distinct words in boxes), result to be among the top influential features that are crucial for AD classifications. Prolonged in-air times are associated with higher SHAP values, suggesting that hesitations or uncertainty during writing are strong indicators of AD.

Similar observations for the other top-listed features can be drawn from Figures 8.3 and 8.4, which summarize and rank the top 20 features of the DARWIN dataset by their importance in predicting AD. Detailed descriptions of these features and the associated tasks are provided in Tables 2.5 and 2.6, respectively.

Figure 8.4 provides additional insight and highlights the influence of feature value on the prediction task. For instance, considering the features, pendown count $num\_pendown19$, performed for task 19 (i.e., while copying the postal order fields), the lower value produces a positive higher SHAP value, hence the greater impact on AD classification. Likely, using Fig 8.3 and Fig 8.4, the relation between the feature value and its impact on AD prediction can be explored. Such an analysis and interpretation can help to have a greater insight into the representative features of AD. Eventually, it helps to improve output while analyzing AD.

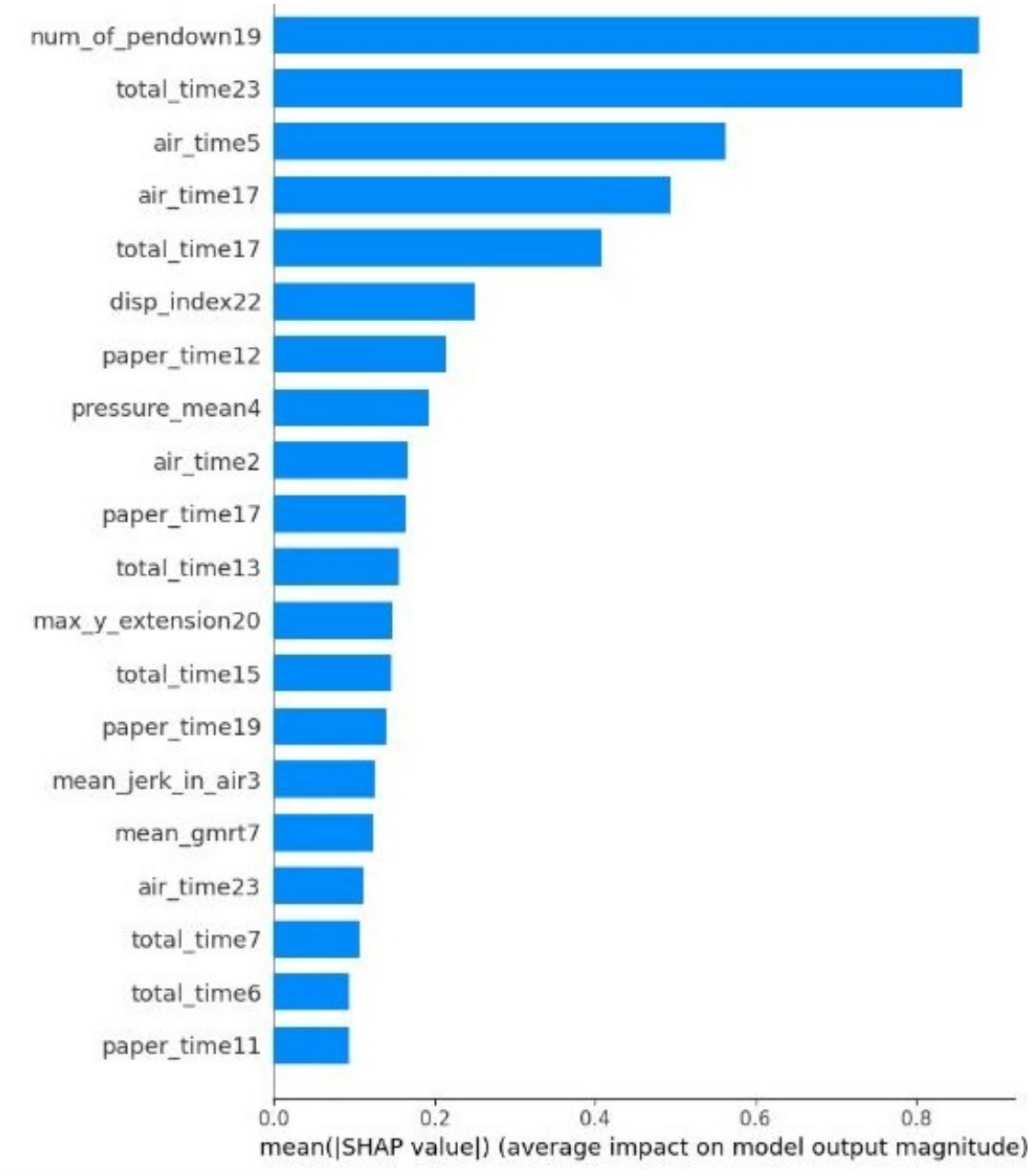

FIGURE 8.3: Feature importance using SHAP

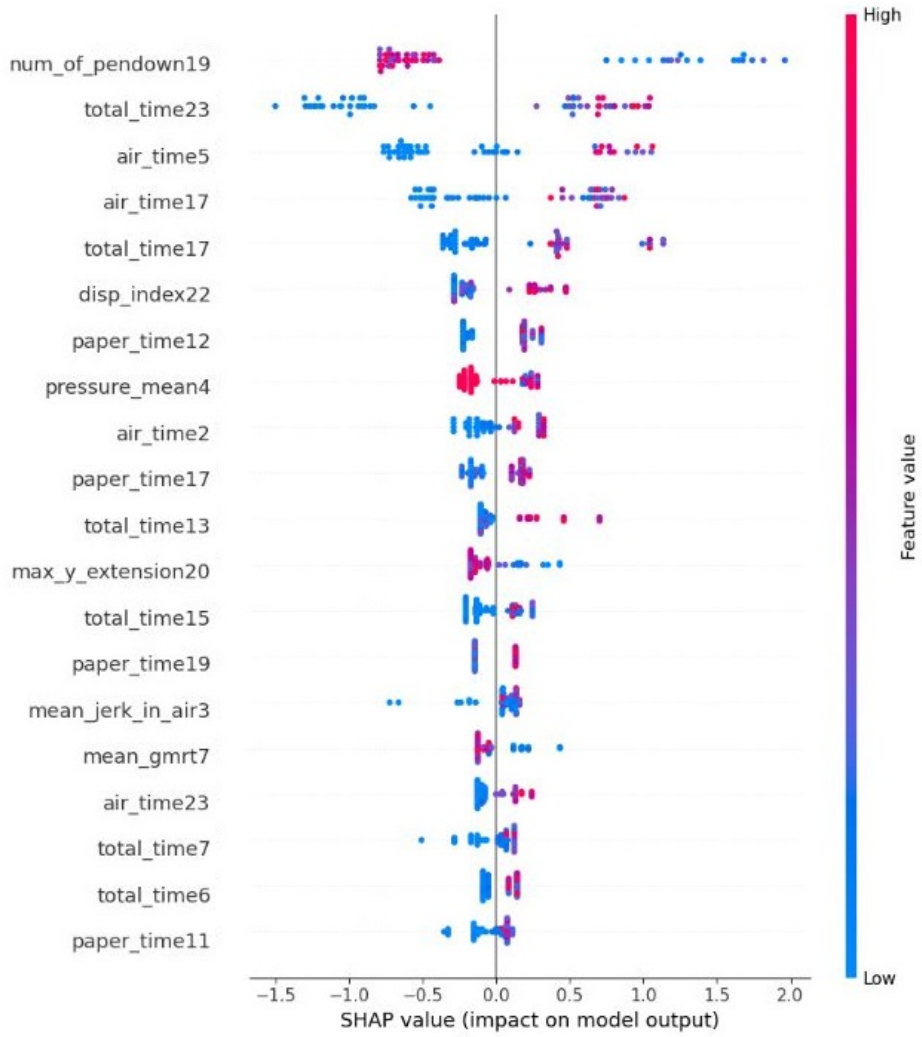

FIGURE 8.4: Summary plot drawn using SHAP

8.3.3 Comparison with state-of-the-art methods

To assess the effectiveness of our Alzheimer's classification approach, we compared it with recent studies using the Darwin dataset. This dataset, recently published, is now a benchmark for evaluating the performance of machine learning models. A quick examination of this comparative study reveals several critical insights. Our proposed model provides performance parameters comparable to those of other recent approaches. The model provides the highest precision of 94.44%. Precision is crucial in the context of medical diagnosis, as false positives can have significant ramifications. Although [314] and [168] exhibit commendable accuracy and precision,

our model surpasses them in terms of precision.Even though our model's accuracy is 88.57

A noteworthy aspect of our proposed model is the emphasis on explainability. Unlike the other referenced methods that do not offer feature importance or model explainability, our model incorporates the SHAP method. This inclusion adds a layer of interpretability, making our model particularly useful in real-world applications where understanding the decision-making process is crucial.

Table 8.2 provides a comprehensive comparison of our proposed model with other contemporary models in the literature, specifically tailored for the DARWIN dataset.

In conclusion, while there are multiple methods tailored for the DARWIN dataset, our proposed stacking approach not only demonstrates competitive performance but also excels in offering model explainability, a feature often overlooked but highly valuable in the domain of medical diagnostics.

TABLE 8.2: Comparison of the state of the art methods on DARWIN dataset

References	Dataset	Approach	Explanability/ Feature Importance	Performance Metrics
[129]	DARWIN	Machine Learning	No	Accuracy: 85.29%
[314]	DARWIN	Machine Learning and PSO algorithm	No	Accuracy: 90.57% Precision: 88.46%
[168]	DARWIN	Deep Learning CNN	No	Accuracy: 90.4% Precision: 92.04%
Proposed model	DARWIN	Machine learning, Stacking	**Yes**	Accuracy: 88.57% **Precision: 94.44%**

8.3.4 Threats to Validity

While the proposed method for Alzheimer's Disease detection shows promise, several factors may affect the validity of the findings. These are outlined below.

Internal Validity

The current evaluation relies solely on the Darwin dataset, which may not reflect the variability seen in broader clinical populations. Using a single dataset risks

introducing data-specific biases, potentially inflating performance. Strengthening internal validity requires testing on multiple datasets to ensure consistent results across different data distributions.

External Validity

The generalizability of the method depends on its performance across diverse patient populations, clinical conditions, and data collection environments. Effective deployment also requires integration into real-world diagnostic workflows. Collaborating with clinicians to validate the system's practical utility will be essential for confirming its relevance and adaptability in clinical settings.

Construct Validity

The model aims to detect Alzheimer's Disease based on features extracted from the dataset, often guided by statistical patterns. However, clinical diagnosis involves complex cognitive, behavioral, and neurological assessments that may not be fully captured by the available features.

Conclusion Validity

The clinical effectiveness of the method cannot be fully determined without real-world validation, including clinical trials and expert feedback. Establishing user trust and ensuring appropriate training are critical for practical adoption. Without these, conclusions about the model's clinical impact may be premature. Future efforts should address these areas to support credible and actionable outcomes.

8.4 Summary

This chapter has presented an innovative approach to utilizing handwriting dynamics, machine learning, and explainable AI for the early detection of Alzheimer's Disease (AD). By focusing on fine motor skills as revealed through handwriting, we

have explored how subtle changes in motion patterns can be indicative of neurological deterioration associated with AD.

Our work involved the use of the DARWIN dataset to examine the correlation between handwriting dynamics and Alzheimer's. The proposed stacking ensemble model has demonstrated its efficacy by achieving a high precision rate of 94.44%, which is crucial in reducing false positives in medical diagnostics. This performance underscores the potential of our approach in providing reliable early-stage predictions of Alzheimer's, thereby improving patient outcomes through timely interventions.

A significant contribution of our study is the integration of SHAP (SHapley Additive exPlanations), which enhances model transparency and interpretability. By identifying and ranking the most influential features, SHAP has allowed us to gain insights into the decision-making process of the model. This transparency is essential in clinical settings, where understanding the rationale behind a diagnosis is as important as the diagnosis itself.

Moving forward, collaboration with medical professionals, including neurologists and geriatricians, will be vital. Their expertise can inform the refinement of our model, ensuring it aligns with clinical needs and practices. Furthermore, real-world validation through clinical trials will be essential to confirm the model's applicability and reliability.