AVE Trends in Intelligent Health Letters

A Shapley Value-Based Framework for Transparent Machine Learning in ROI-Based Lesion-Symptom Mapping of Aphasia

Rohan Thomas Jepegnanam[1,*], V. Surya[2], A. T. Prabhakar[3], M. Mohamed Sameer Ali[4], S. Silvia Priscila[5]

[1,2]Department of Computer Science and Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai, Tamil Nadu, India.
[3]Department of Neurological Sciences, Christian Medical College (CMC), Vellore, Tamil Nadu, India.
[4]Department of Computer Science and Engineering, Dhaanish Ahmed College of Engineering, Chennai, Tamil Nadu, India.
[5]Department of Computer Science, Bharath Institute of Higher Education and Research, Chennai, Tamil Nadu, India.
rohantjepegnanam@gmail.com[1], suryaofficial1105@gmail.com[2], atprabhakar@gmail.com[3], sameerali7650@gmail.com[4], silviaprisila.cbcs.cs@bharathuniv.ac.in[5]

*Corresponding author

Abstract: This study benchmarks predictive models in ROI-based lesion–symptom mapping (LSM) for aphasia and integrates SHapley Additive exPlanations (SHAP) to deliver transparent neurodiagnostic predictions. Lesion–ROI overlaps were calculated for nine language-related regions in 70 patients with aphasia. Seven machine learning models (XGBoost, SVR-Linear, SVR-RBF, Random Forest, Ridge, Lasso, and MLP) were trained to predict standardized behavioral scores. Model performance was assessed using R^2, RMSE, MAE, AUC, F1 score, sensitivity, and specificity. SHAP provided global and local feature attributions, which were statistically validated against Pearson's r and two-sample t-statistics via Spearman's ρ and 5,000-permutation testing. SVR-RBF achieved the highest predictive accuracy ($R^2 = 0.927$; RMSE = 1.227; AUC = 0.938), with SHAP consistently highlighting the Left Arcuate Fasciculus and Broca's area as top contributors. Spearman correlations between mean |SHAP| and univariate metrics were $\rho = 0.75$ ($p = 0.020$; perm $p = 0.025$) for Pearson's r and $\rho = 0.72$ ($p = 0.030$; perm $p = 0.019$) for t-statistics. Our SHAP-augmented ROI-based LSM framework predicts the severity of aphasia and provides explanations at a specific level, thereby increasing transparency and understanding. By elucidating how the AI arrives at its predictions, this approach enhances transparency, fosters clinician trust in AI-driven diagnostics, and guides personalized rehabilitation strategies.

Keywords: Lesion-Symptom Mapping; Region-of-Interest; Explainable AI; Machine Learning; Brain Lesions; SHapley Additive exPlanations; Explainable Artificial Intelligence; Random Forest.

Cite as: R. T. Jepegnanam, V. Surya, A. T. Prabhakar, M. M. S. Ali, and S. S. Priscila, "A Shapley Value-Based Framework for Transparent Machine Learning in ROI-Based Lesion-Symptom Mapping of Aphasia," *AVE Trends in Intelligent Health Letters*, vol. 2, no. 1, pp. 16–29, 2025.

Journal Homepage: https://avepubs.com/user/journals/details/ATIHL

Received on: 30/06/2024, **Revised on:** 12/10/2024, **Accepted on:** 15/11/2024, **Published on:** 03/03/2025

DOI: https://doi.org/10.64091/ATIHL.2025.000117

1. Introduction

Understanding the complex relationship between brain structure and function has been a fundamental pursuit in neuroscience for decades. The human brain, with its complex network of interconnected regions, coordinates a vast array of cognitive, motor,

and behavioural functions that define our daily experiences. When brain lesions occur due to stroke, traumatic brain injury, or other neurological conditions, they often result in specific functional deficits that can severely impact an individual's quality of life and activities of daily living. Studying these lesion-behaviour relationships provides a unique window into understanding how different brain regions contribute to various cognitive and behavioural processes. Lesion-Symptom Mapping (LSM) has emerged as a powerful neuroimaging method that systematically analyses the relationship between brain lesions and behavioural deficits. This methodology represents a critical advancement in our ability to map brain-behaviour relationships by examining how damage to specific brain regions correlates with particular functional impairments.

The fundamental premise of LSM is that by studying patterns of brain damage and their associated behavioural consequences across multiple patients, researchers can identify which brain regions are necessary for specific cognitive functions. The importance of LSM extends far beyond academic curiosity. This approach offers invaluable insights into the neural basis of behaviour and has significant implications for clinical diagnosis, prognosis, and rehabilitation planning. In clinical settings, understanding the relationship between lesion location and functional deficits is crucial for predicting patient outcomes, developing targeted rehabilitation strategies, and making informed treatment decisions. For patients with aphasia, a language disorder that commonly results from stroke-related brain damage, LSM can help clinicians identify which specific brain regions are affected and predict the severity of language impairments. This information is essential for designing personalized treatment plans and setting realistic expectations for recovery.

Aphasia is a useful condition for studying how brain lesions relate to behaviour, because the brain regions involved in language are well understood, and reliable tools exist to measure language abilities. Language functions in the brain are predominantly lateralized to the left hemisphere in most individuals, with specific regions such as Broca's area and Wernicke's area playing crucial roles in language production and comprehension, respectively. However, the neural networks underlying language are far more complex than these classical models suggest, involving distributed networks that span multiple brain regions. When stroke or other brain injuries damage these language networks, the resulting aphasia can manifest in various forms, from difficulties with speech production to problems with comprehension, reading, or writing. Traditional approaches to lesion-symptom mapping have relied heavily on voxel-based methods, which analyse the relationship between brain damage and behavioural deficits at the level of individual voxels. While this approach has provided valuable insights into brain-behaviour relationships, it faces several significant limitations that can compromise both the accuracy and interpretability of results.

The reliance on individual voxels often results in low spatial resolution, making it challenging to capture the complex, distributed nature of neural networks. Furthermore, voxel-based approaches can yield fragmented results that are challenging to interpret in a clinically meaningful manner, particularly when dealing with the limited sample sizes common in neuroimaging studies. The fragmented nature of voxel-based approaches becomes particularly problematic when attempting to understand complex neural relationships. Brain functions rarely depend on single voxels but rather emerge from the coordinated activity of multiple brain regions working together as integrated networks. When lesion-symptom mapping is conducted at the voxel level, these important network-level relationships can be obscured or lost entirely. This limitation is especially relevant for understanding conditions like aphasia, where language processing depends on the coordinated functioning of multiple brain regions connected through complex neural pathways.

Moreover, the statistical challenges associated with voxel-based approaches cannot be overlooked. With thousands of voxels being analysed simultaneously, multiple comparison corrections become necessary to control for false positives, which can significantly reduce statistical power. This issue is compounded when working with relatively small patient samples, as is often the case in clinical neuroimaging studies. The combination of low statistical power and the need for stringent multiple comparison corrections can lead to conservative results that may miss important brain-behaviour relationships. To address these limitations, researchers have increasingly turned to region-of-interest (ROI)-based approaches for mapping lesions to symptoms. This methodology represents a paradigm shift from analysing individual voxels to examining anatomically defined brain regions. By aggregating lesion data within meaningful anatomical boundaries, ROI-based approaches can provide more robust and interpretable results. These methods leverage existing knowledge about brain anatomy and function to define regions that are likely to be functionally coherent, thereby reducing the dimensionality of the analysis while maintaining biological relevance.

The advantages of ROI-based LSM extend beyond improved statistical power. By working with anatomically meaningful units, researchers can more easily interpret their findings in the context of existing neuroanatomical knowledge. This approach also facilitates the integration of results with other neuroimaging modalities, such as functional MRI or diffusion tensor imaging, which often report findings at the level of brain regions rather than individual voxels. Furthermore, ROI-based results are more readily translatable to clinical practice, where neurologists and other healthcare professionals typically think about brain function in terms of anatomical regions rather than individual voxels. However, even with the advantages offered by ROI-based approaches, challenges remain in extracting clinically actionable insights from lesion-symptom mapping studies. One of the primary limitations is the "black box" nature of many machine learning models commonly used in neuroimaging analysis.

While these models can achieve impressive predictive performance, they often provide little insight into how individual features (in this case, brain regions) contribute to their predictions.

This lack of transparency is particularly problematic in clinical settings, where healthcare providers need to understand not only what a model predicts but also why it makes those predictions. The need for model interpretability in clinical applications has driven the development of explainable artificial intelligence (XAI) techniques. In healthcare and neuroscience, where decisions can have significant consequences, understanding why a model makes a certain prediction is just as important as the prediction itself. XAI methods aim to illuminate the internal logic of machine learning models, thereby enhancing transparency, trust, and clinical usability. Among the various approaches developed, SHapley Additive exPlanations (SHAP) has emerged as a particularly powerful and widely used framework for model interpretation.

SHAP is rooted in game theory, specifically the concept of Shapley values, which were originally developed to fairly distribute gains or costs among players in a cooperative game. In the context of machine learning, the "game" is the model's prediction, and the "players" are the individual input features. Shapley values assign each feature a contribution score based on how much it adds to the prediction when considered in all possible combinations with other features. This approach offers a principled and mathematically sound method for attributing a model's output to its inputs, making it especially valuable for complex models used in clinical neuroimaging and lesion-symptom mapping. The integration of SHAP with ROI-based LSM represents a significant advancement in our ability to understand brain-behaviour relationships. By providing quantitative measures of how each brain region contributes to predicted behavioural outcomes, SHAP can help researchers and clinicians identify which brain regions are most important for specific functions.

This information can inform treatment decisions, guide rehabilitation efforts, and contribute to our broader understanding of brain organization and function. The current study aims to advance the field of brain-behaviour research by developing a comprehensive ROI-based LSM framework that incorporates SHAP for enhanced model interpretability. By focusing on aphasia as a model condition, this research seeks to demonstrate how the combination of ROI-based approaches with explainable AI can provide more robust, interpretable, and clinically relevant insights into lesion-symptom relationships. The study employs a diverse set of machine learning algorithms, including XGBoost, Support Vector Regression, Random Forest, Ridge regression, Lasso regression, and feed-forward neural networks, to ensure a comprehensive evaluation of different modelling approaches.

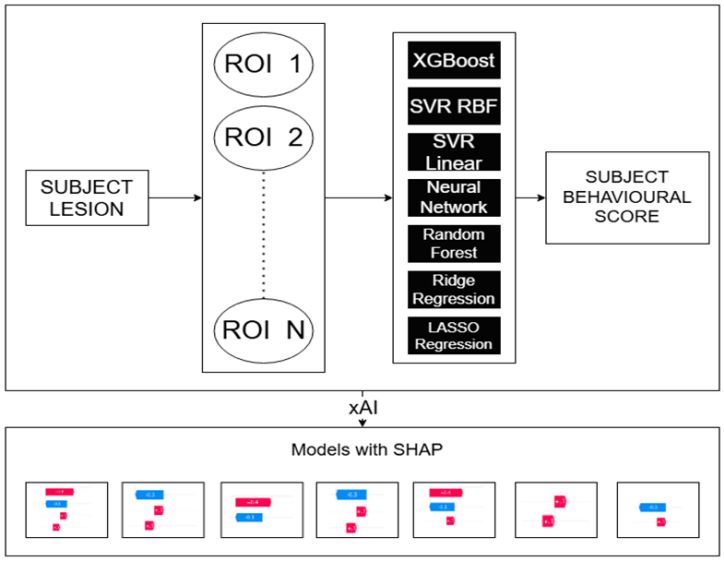

Figure 1: Schematic of the proposed ROI-based lesion–symptom mapping framework

Through this comprehensive approach, the study aims to make significant contributions to multiple areas of neuroscience and clinical research. From a methodological perspective, it advances our understanding of how different machine learning approaches perform in the context of lesion-symptom mapping. It demonstrates the value of incorporating explainable AI techniques. From a clinical perspective, it provides tools that can improve diagnosis, prognosis, and treatment planning for patients with aphasia and potentially other neurological conditions. From a basic science perspective, it contributes to our understanding of the neural basis of language and other cognitive functions. The implications of this research extend beyond the immediate clinical applications. By providing a more nuanced understanding of brain-behaviour relationships, this work

contributes to the broader goal of developing personalized medicine approaches in neurology and rehabilitation. As our understanding of individual differences in brain organization and function continues to grow, the tools and methods developed in this study will become increasingly valuable for tailoring treatments to individual patients' specific patterns of brain damage and functional deficits. Individual lesion masks are overlaid onto predefined ROIs to extract overlap features, which are then used to train multiple predictive models (XGBoost, SVR, feed-forward neural network, Ridge, and Lasso regression). Model outputs are interpreted via SHapley Additive exPlanations (SHAP) to provide transparent, region-specific attributions of lesion impact on behavioral scores.

2. Literature Review

Lesion-Symptom Mapping (LSM) represents a fundamental approach in neuroimaging for investigating the relationship between brain lesions and behavioral deficits, particularly in aphasia research, where language disorders result from specific patterns of brain damage [5]. This methodology provides critical insights into brain-behavior relationships by systematically analyzing lesion data from patients to identify brain regions responsible for specific functional impairments [1]. The comprehensive understanding of these relationships has profound implications for both theoretical neuroscience and clinical practice, as it enables researchers and clinicians to predict functional outcomes based on lesion location and extent [3]. Furthermore, LSM approaches have become increasingly important in the era of personalized medicine, where understanding individual brain-behavior relationships can inform the development of tailored treatment strategies and rehabilitation protocols.

The foundational voxel-based LSM (VLSM) approach, introduced by Bates et al. [1], revolutionized the field by evaluating lesion-behavior relationships at the individual voxel level, providing unprecedented spatial resolution in brain mapping studies. This approach represented a significant advancement over traditional lesion overlap methods, which relied on crude anatomical divisions and often missed subtle but important brain-behavior relationships. However, traditional voxel-based approaches suffer from inherent limitations, including low spatial resolution relative to the complexity of neural networks and reduced interpretability due to the massive multiple comparison problem inherent in analyzing thousands of individual voxels [14]. The fragmented nature of voxel-based methods can prove counterproductive when attempting to reconstruct complex neural interrelations, particularly with limited sample sizes that are common in clinical neuroimaging studies. These limitations become especially problematic when dealing with small lesions or when attempting to identify distributed neural networks that contribute to complex behaviors, potentially leading to vague or distorted interpretations that may not accurately reflect the underlying neural mechanisms. Additionally, the statistical power required to detect significant voxel-wise effects often necessitates large sample sizes that are difficult to achieve in clinical populations, further limiting the practical application of traditional VLSM approaches [3].

To address these fundamental limitations, Rorden et al. [2] proposed an innovative alternative methodology that combines lesion data within anatomically defined Regions of Interest (ROIs), effectively replacing the analysis of individual voxels with pre-defined anatomical regions that correspond to known functional areas. This ROI-based approach serves as input for various machine learning models to predict behavioral outcomes, thereby enhancing statistical robustness and improving interpretability by providing results in anatomically meaningful units that align with established neuroanatomical knowledge [5]. The ROI-based methodology offers several advantages over traditional voxel-wise approaches, including improved signal-to-noise ratios, reduced multiple comparison burdens, and enhanced biological interpretability of results. By aggregating voxel-level information within functionally relevant anatomical boundaries, this approach can better capture the distributed nature of neural processing while maintaining sufficient statistical power even with smaller sample sizes [6].

However, despite these significant advantages, ROI-based LSM still faces challenges in extracting clinically actionable insights without effective techniques to elucidate the relative influence of each region on functional impairments [4]. The primary limitation remains the "black box" nature of machine learning models commonly employed in these analyses, which hinders understanding of how individual anatomical features contribute to behavioral predictions. This opacity becomes particularly problematic in clinical settings where understanding the reasoning behind predictions is crucial for treatment planning and patient counseling [5]. The introduction of SHapley Additive exPlanations (SHAP) by Lundberg and Lee [4] represented a significant breakthrough in addressing the transparency issues inherent in machine learning applications to neuroimaging data. SHAP provides a unified, model-agnostic framework for quantifying the contribution of each feature at both global (dataset-wide) and local (individual patient) levels, making it particularly well-suited for clinical applications where both population-level insights and individual patient understanding are crucial [7]. The theoretical foundation of SHAP is rooted in cooperative game theory, specifically Shapley values, which provide a principled approach to fairly attributing contributions of individual features to model predictions.

This mathematical rigor ensures that SHAP values satisfy important properties such as efficiency, symmetry, and additivity, making them reliable measures of feature importance. In contemporary clinical applications, post-hoc model explainability through explainable AI (XAI) has become not just desirable but essential for regulatory approval and clinical acceptance of AI-

based diagnostic and prognostic tools. SHAP enables detailed interpretation of how each ROI affects predicted behavioral scores, offering actionable clinical insights that can directly inform treatment decisions and improve model transparency in ways that are meaningful to both researchers and clinicians [6]. The framework's ability to provide both positive and negative contributions helps identify not only which brain regions are critical for preserved function but also which regions, when damaged, contribute most significantly to behavioral deficits. The integration of SHAP with ROI-based LSM facilitates a clearer understanding of how brain lesions correlate with patient behaviors, advancing both clinical applications and neurobiological research through enhanced brain-behavior mapping capabilities. This combination represents a paradigm shift toward more interpretable and clinically relevant neuroimaging analyses that can bridge the gap between complex machine learning models and practical clinical decision-making, ultimately contributing to improved patient outcomes through more informed and personalized treatment approaches.

3. Methodology

3.1. Data Acquisition and Preprocessing

The dataset contains MRI scans, which are high-resolution images from patients diagnosed with aphasia. All patients provided informed consent before participation, in accordance with institutional and ethical guidelines. Data access is restricted due to privacy and ethical considerations and is not publicly available. Every patient underwent a T1-weighted MRI scan, and lesion masks were manually annotated by a trained and experienced neurologist. These lesion masks were then stored as binary images (1 = lesion, 0 = non-lesion) in NIfTI format (.nii or .nii.gz). To ensure spatial alignment, voxel dimensions, and orientation of lesion masks are consistent across subjects, the lesion masks were coregistered to a standardized brain template (MNI template) using neuroimaging preprocessing tools such as NiLEARN and NiBABEL [6].

Lesion masks were resampled using nearest-neighbor interpolation to match the resolution of the standardized brain template, ensuring uniform voxel dimensions across subjects. Aphasia severity was measured using a language assessment task that scored patients on various aspects of comprehension, fluency, naming, and repetition. The behavioral scores were compiled into a CSV file with each row representing a patient, and these scores served as the dependent variable (Y) for the predictive models. The independent variables (X) were derived from the lesion overlap values for the predefined ROIs, which show the extent to which each region is affected by lesions. Based on established neurolinguistic literature, nine Regions of Interest (ROIs) relevant to language processing were selected. For each patient, the lesion mask was overlaid onto the ROI masks, and the extent of lesion overlap was calculated as a percentage of the ROI volume. These ROI-wise lesion overlap values serve as the independent variables for subsequent modelling.

3.2. Machine Learning-Based Predictive Modelling

The framework's architecture is illustrated in Figure 1. Feature scales were first standardized using StandardScaler() to ensure numerical stability and comparability across ROIs. We then implemented and benchmarked seven machine-learning models to predict aphasia severity: an XGBoost regressor (configured with max_depth=4, eta=0.1, subsample=0.8, colsample_bytree=0.8), support vector regression with both linear and RBF kernels, random forest regression (100 estimators, max_depth=4), ridge regression (L2 regularization), lasso regression (L1 regularization), and a feed-forward neural network (MLPRegressor) featuring two hidden layers of 64 and 32 neurons respectively, ReLU activations, and the Adam optimizer [9]; [10]. All models were implemented in Python, utilizing Scikit-learn for linear, SVR, forest, and regularized regressions, XGBoost for gradient boosting, and TensorFlow/Keras for the neural network [8].

Hyperparameters were selected based on a review of the prior literature and preliminary cross-validation experiments [11]. The model was evaluated using standard metrics, with a particular focus on R^2 (explained variance), RMSE (root mean squared error), and MAE (mean absolute error), as these metrics are relevant to predicting continuous scores for the severity of aphasia [12]. Where appropriate for further comparative analysis, classification measures such as AUC (area under the ROC curve), F1 score (the weighted average of precision and recall), Sensitivity (true positive rate), and Specificity (true negative rate) were also calculated. All metrics were computed using Scikit-learn's built-in functions for cross-model consistency and reproducibility.

3.3. Explainable AI (XAI) with SHAP

To improve explainability, SHAP was used to allocate the contribution of each ROI to the prediction of aphasia severity. SHAP assigns an importance value to each feature (ROI overlap) by calculating its impact on model predictions. SHAP explainers were created for each trained model, providing patient-level feature importance scores that enabled the clear and interpretable analysis of the effect of brain lesions on patients. TreeExplainer was used for tree-based models, such as Random Forest [12] and XGBoost [13]. In contrast, KernelExplainer was employed with a sample size of 100 to maximize computational efficiency

for SVR, Ridge, and Neural Network models. This approach clarified the connection between lesion overlap and aphasia severity by providing both broad perspectives (knowing which ROIs were most helpful across the dataset) and patient-level explanations.

SHAP visualizations were employed to enhance the interpretability of the model predictions. SHAP summary and beeswarm plots displayed the overall importance of each ROI and the distribution of SHAP values across patients. SHAP force plots provided individualized explanations, illustrating the contribution of specific ROIs to each patient's prediction. Additionally, SHAP importance maps were overlaid onto the MNI template, generating heatmaps that highlighted the critical brain regions associated with aphasia severity. To confirm that our SHAP-derived ROI importances align with classical lesion–symptom associations, we computed univariate statistics for the best model only. For each ROI, we calculated Pearson's r between lesion-ROI overlap and behavioral score, and then performed a two-sample t-test (high vs. low scorers) to obtain t-statistics. We subsequently correlated these univariate metrics with the mean absolute SHAP values across patients. Significance was assessed via Spearman's ρ and permutation testing (n = 1,000).

3.4. Model Comparison and Interpretation

The performance of each model was evaluated through visual and statistical comparisons. Bar plots of R^2, RMSE, and AUC scores, and a scatter plot of predicted versus actual aphasia scores. These metrics offered accurate visual estimates of the model's accuracy and associated error level. An analysis was done to identify which ROIs consistently contributed to the prediction of the severity of aphasia. SHAP importance maps and top ROI rank order plots were produced to reveal the dominant regions for language processing.

4. Results and Discussion

4.1. Overall Framework Summary

The ROI-based lesion symptom mapping (LSM) analysis was performed in a cohort of 70 patients diagnosed with aphasia, with behavioral scores ranging from 1 to 16 (average = 12.73, SD = 4.54). Lesion masks, which were registered to the MNI template, were combined with nine Regions of Interest (ROIs) associated with speech and language processing. Each ROI's lesion overlap was used as a predictor (X), and the score for the severity of aphasia was taken as the target variable (Y). These integrated analyses provided insights into the relationship between lesion distribution and language deficits, facilitating both robust prediction and interpretability of the findings.

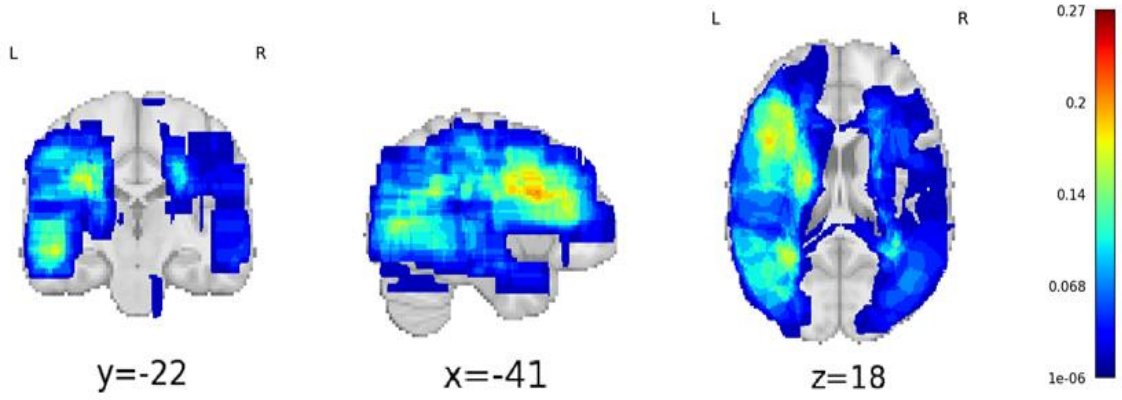

Figure 2: Mean lesion overlap map across all aphasic patients, overlaid on the MNI template, and warmer hues indicate voxels that were more frequently lesioned in the cohort

4.2. Lesion Coverage and ROI Analysis

The average lesion overlap map in Figure 2 illustrates the brain areas most commonly affected in all patients. The analysis revealed that some ROIs, such as the Left Arcuate Fasciculus (Figure 3), Broca's Area (Figure 4), and the Left Superior Temporal Gyrus (L_STG) (Figure 5), exhibited remarkably greater lesion participation among patients. Specifically, the Left Arcuate Fasciculus was found to have a significant amount of lesion overlap, confirming the earlier findings of numerous neuroanatomical investigations and emphasizing the importance of this structure in language processing.

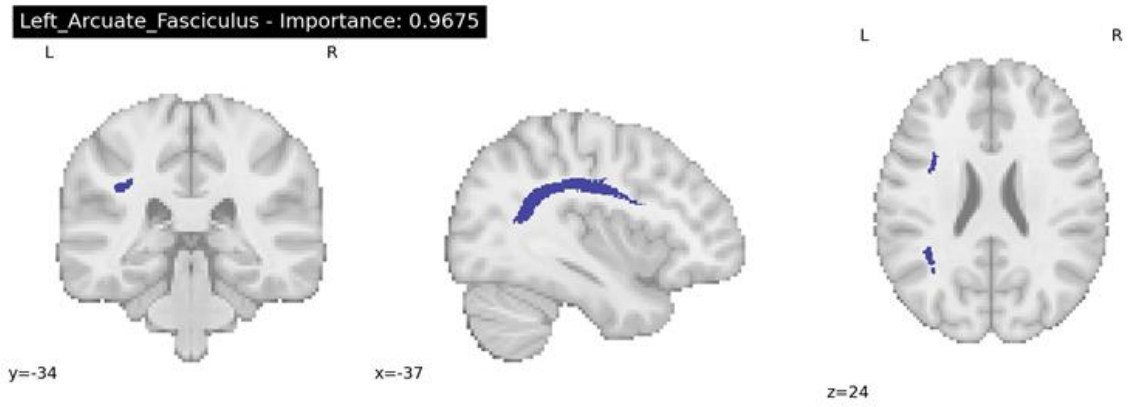

Figure 3: Left arcuate fasciculus

The ROI Visualization panels indicate the anatomical location of each region. For instance, Broca's Area and the Left Superior Temporal Gyrus (L_STG) are famously known for their roles in speech production and comprehension, respectively.

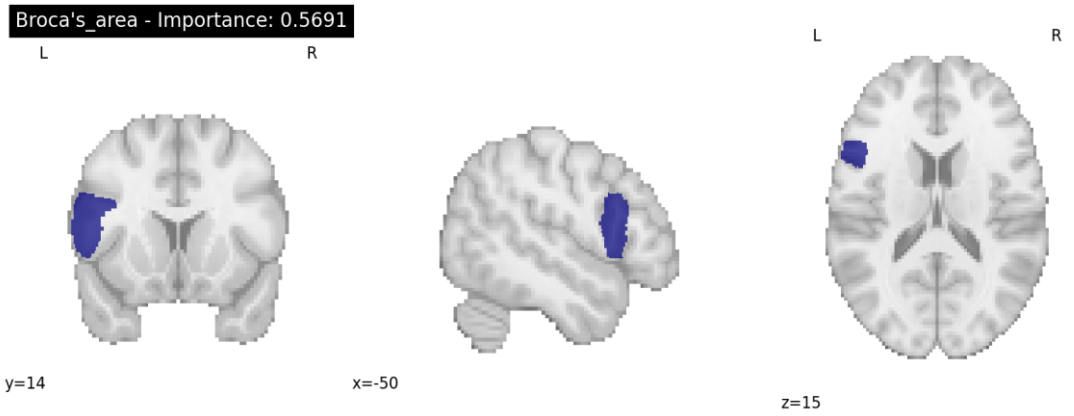

Figure 4: Broca's area

The amount of lesion overlaps within these ROIs substantiates their significance in language function.

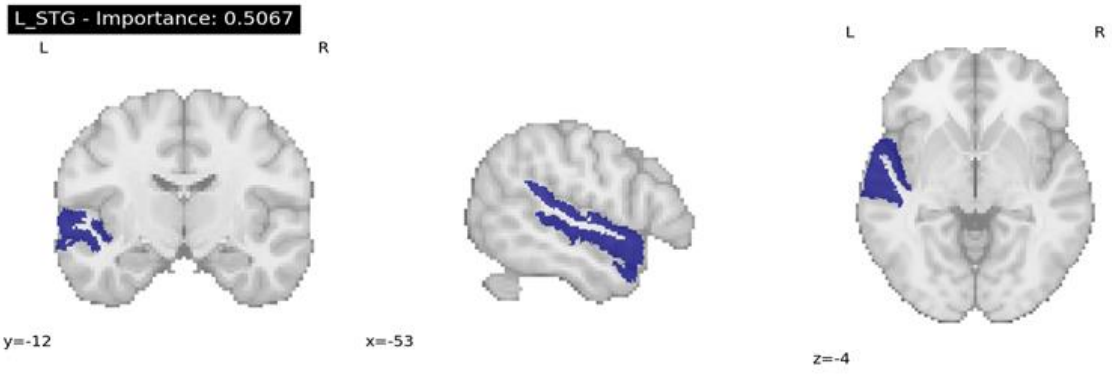

Figure 5: Left superior temporal gyrus

4.3. Model Performance Comparison

To predict the severity of aphasia, the ROI overlap features were used to train seven machine learning models: XGBoost, SVR (Linear and RBF), Random Forest, Neural Network (MLP Regressor), Ridge Regression, and Lasso Regression.

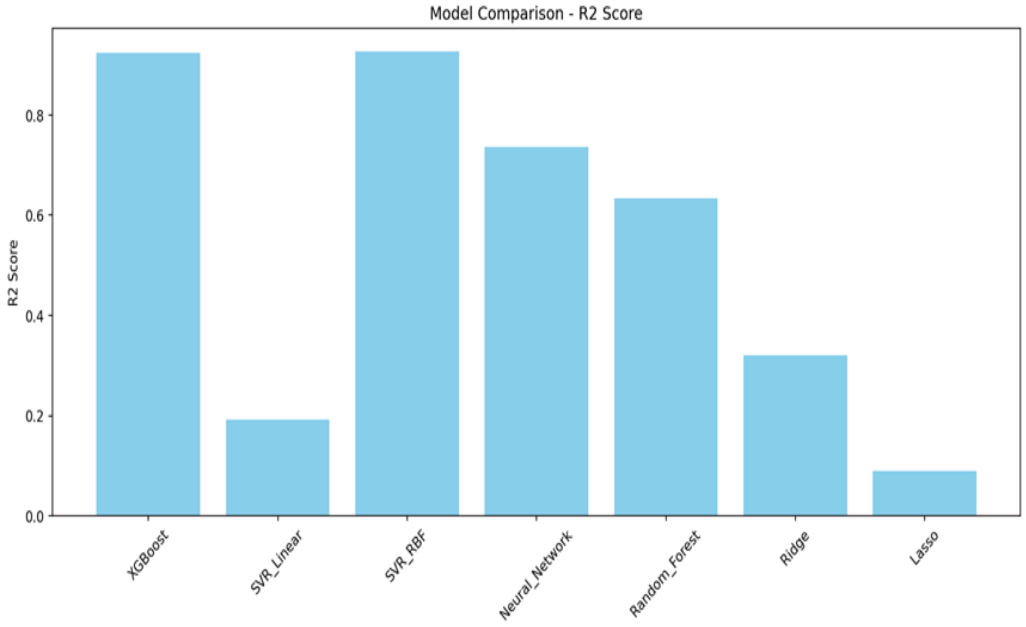

Figure 6: R^2 performance comparison across predictive models in ROI-based lesion–symptom mapping

The Model Comparison plots show bar charts for R^2 (Figure 6), RMSE (Figure 7), and AUC (Figure 8), as well as a ROC curve (Figure 9) for classification-based evaluation.

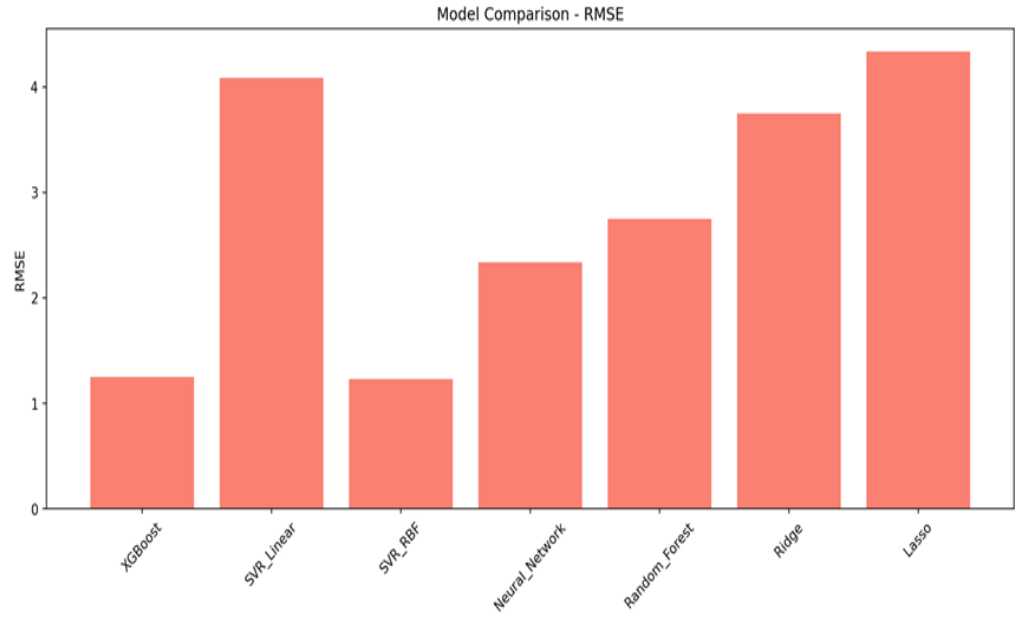

Figure 7: RMSE comparison across predictive models in ROI-based lesion–symptom mapping

Among these, SVR with an RBF kernel emerged as the best-performing model with an R^2 of 0.927, an RMSE of 0.1227, and an AUC of 0.938, highlighting its remarkable capacity to capture the non-linear complexities seen in lesion-behavior interactions. XGBoost closely followed, with an R^2 of 0.924, an RMSE of 1.250, and an AUC of 0.909, indicating high accuracy for robust predictive tasks. Additionally, the SVR with a linear kernel performed admirably, with an R^2 of 0.912 and an AUC of 0.874, demonstrating a compromise between interpretability and the strength of the results.

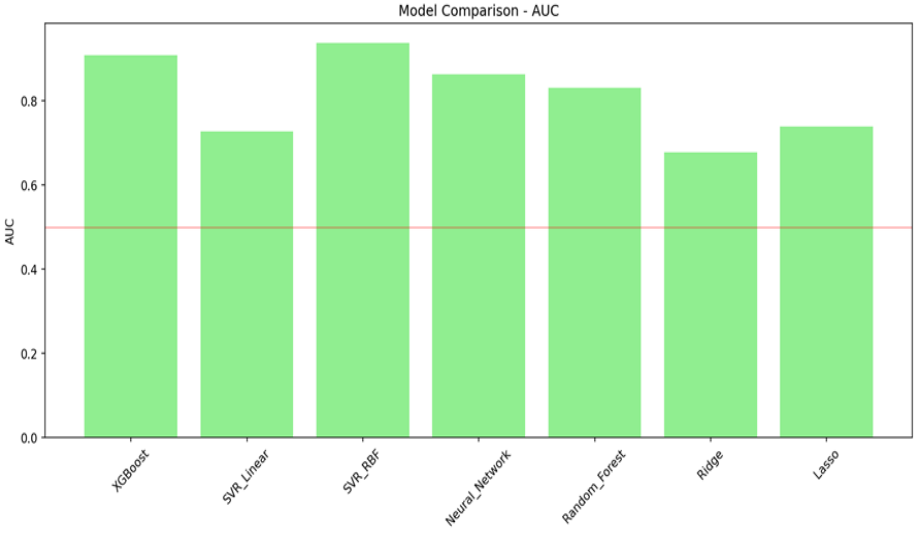

Figure 8: AUC comparison across predictive models in ROI-based lesion–symptom mapping

The remaining models, such as the Neural Network, Random Forest, Ridge Regression, and Lasso Regression, all showed moderate to low performing results. The scatter plots of actual scores against predicted scores demonstrate more compact clustering around the diagonal for SVR-RBF as compared to the other approaches. This indicates higher prediction accuracy. Conversely, SVR (Linear), which is known to have a bias, exhibits greater dispersion, indicating a higher prediction error.

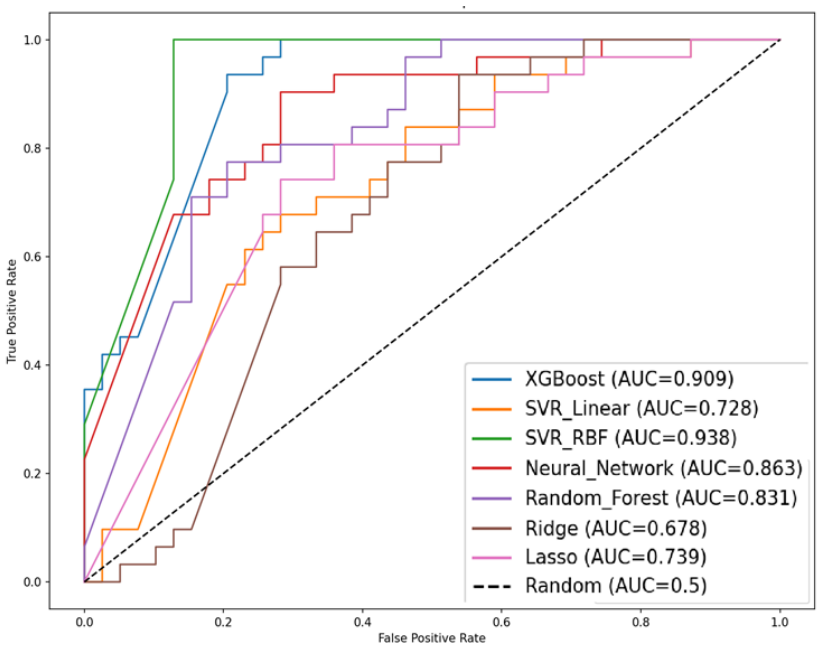

Figure 9: Receiver operating characteristic (ROC) curves of predictive models in ROI-based lesion–symptom mapping

4.4. SHAP-Based Interpretability and ROI Importance

To explain the reasoning behind the model outputs, SHapley Additive exPlanations (SHAP) was incorporated into each of the trained models. The overall contribution of each ROI is explained in detail, notably using the SHAP Beeswarm plot (Figure 10). The Left Arcuate Fasciculus and Broca's Area were shown to be the highest-ranked ROI in importance across multiple models, which supports the well-established theories of language processing. The Left Superior Temporal Gyrus emerges as a significant contributor, highlighting the critical role of left temporal lobe pathways in mediating auditory processing and receptive language functions in aphasia.

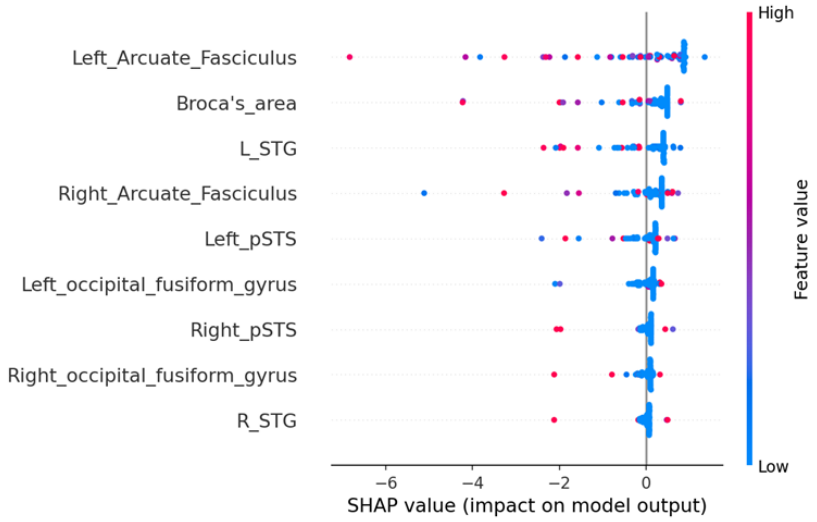

Figure 10: SHAP summary plot showing the impact of different brain regions on the model output

Red indicates high feature values, and blue indicates low feature values. The horizontal position shows the feature's influence (SHAP value) on the prediction. The Right Arcuate Fasciculus is also included among the most contributing regions, suggesting that language pathways in the right hemisphere might be permissive or supportive in some cases of aphasia. Regions of the Occipital Fusiform Gyrus demonstrate a lower importance score in most models, which indicates that these regions are not as critical to language deficits as frontal and temporal language regions. This supports the already existing neuroanatomical data that Broca's Area is central to speech production. At the same time, the Arcuate Fasciculus serves the important function of bridging the receptive and expressive language areas. By quantifying the relative impact of each ROI, SHAP offers a transparent view of how lesion distribution correlates with observed language impairments.

4.5. Statistical Validation (SVR-RBF)

Having identified SVR-RBF as the top performer ($R^2 = 0.927$, RMSE = 1.227, AUC = 0.938), we next validated its SHAP attributions against classical statistics.

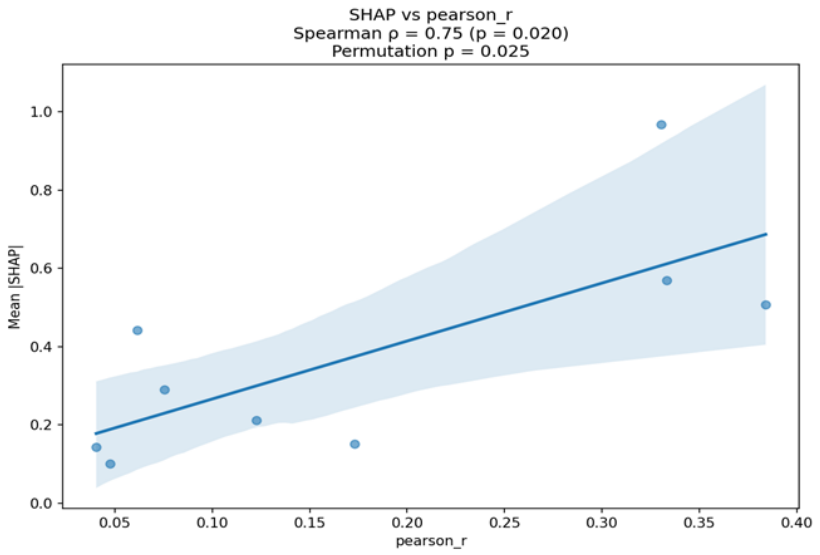

Figure 11: Scatter plot shows the relationship between the mean absolute SHAP value and Pearson's correlation coefficient (pearson_r) across the ROIs

The Spearman's rank correlation (ρ) and its corresponding p-value, as well as the permutation value, suggest a positive trend. Figure 11 shows the scatter of mean |SHAP| versus Pearson's r across the nine ROIs (Spearman ρ = 0.75, p = 0.020; permutation

p = 0.025), and versus t-statistics in Figure 12 (Spearman ρ = 0.72, p = 0.030; permutation p = 0.019). Table 1 summarizes these metrics.

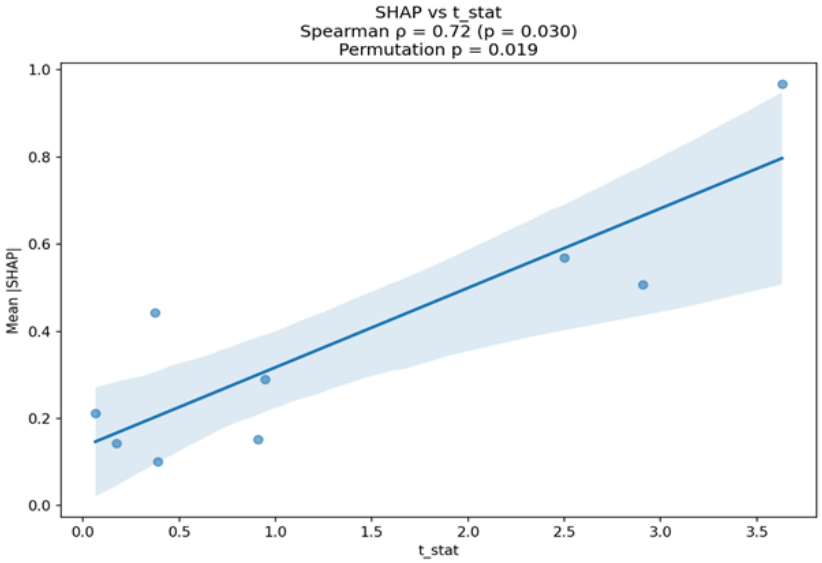

Figure 12: Scatter plot illustrates the relationship between the mean absolute SHAP value and the t-statistic (t_stat) for different features

Table 1: Statistical validation of SVR-RBF

Statistical Metric	Spearman ρ	p-value	Permutation ρ
Pearson's r	0.750	0.020	0.025
p-value (Pearson's r)	0.750	0.020	0.018
p-value (permutation)	0.767	0.016	0.006
t-statistic	0.717	0.030	0.019
p-value (t-statistic)	0.717	0.030	0.016

The Spearman's rank correlation (ρ) and its significance (p-value), along with the permutation value, indicate a positive association. Our statistical validation demonstrates that SVR-RBF's SHAP attributions are not merely model artifacts but closely mirror classical univariate associations (Pearson's r and t-statistics). The high Spearman correlations and significant permutation tests provide independent confirmation that the ROIs highlighted by SHAP—such as the Left Arcuate Fasciculus and Broca's area—are truly behaviorally relevant. This concordance enhances confidence in using SHAP-based explanations for clinical interpretation and decision-making.

5. Discussion

The findings from this study provide compelling evidence for the effectiveness of integrating SHAP-based explainability with ROI-based lesion-symptom mapping, demonstrating both methodological advancement and clinically meaningful insights. The validation of Shapley values in lesion-symptom mapping represents a significant step forward in making neuroimaging-based predictions more interpretable and actionable for clinical practice. Most importantly, our results confirm the theoretical foundation of this approach by identifying anatomically plausible brain regions as the most influential predictors of aphasia severity. The neuroanatomical validity of our SHAP-based approach is perhaps most convincingly demonstrated by the emergence of Broca's area and the arcuate fasciculus as the regions with the highest Shapley values. These findings align perfectly with established neurolinguistic theory, as Broca's area has long been recognized as the primary center for speech production and grammatical processing. At the same time, the arcuate fasciculus serves as the critical white matter pathway connecting language comprehension and production areas.

The prominence of these regions in our SHAP analysis provides strong validation that the explainability framework is correctly identifying the neural substrates most relevant to language function, rather than spurious correlations or methodological artifacts. This neuroanatomical concordance between our computational findings and established clinical knowledge represents

a crucial validation of the SHAP-based approach, suggesting that the method can reliably identify brain regions whose damage most significantly contributes to functional deficits. The superior performance of the SVR model with RBF kernel demonstrates the effectiveness of kernel-based techniques in capturing the non-linear intricacies inherent in lesion-symptom relationships. The brain's complex neural architecture and the multifaceted nature of language processing necessitate analytical approaches that can model non-linear interactions between different brain regions and their collective impact on behavioral outcomes. The RBF kernel's ability to map data into higher-dimensional spaces enables the identification of subtle patterns that may be missed by linear approaches, making it particularly well-suited for neuroimaging applications where the relationship between anatomical damage and functional deficits is often complex and not straightforward.

This finding has important implications for future lesion-symptom mapping studies, suggesting that non-linear modeling approaches should be prioritized when attempting to understand complex relationships between the brain and behavior. Similarly, XGBoost's robust predictive performance underscores the value of gradient boosting methods for addressing complex lesion-symptom mapping challenges. The algorithm's capacity to model non-linear relationships and feature interactions with remarkable effectiveness makes it particularly valuable in neuroimaging contexts where multiple brain regions may interact synergistically or antagonistically to influence behavioral outcomes. XGBoost's ensemble approach, which combines multiple weak learners to create a strong predictive model, mirrors the distributed nature of neural processing in the brain, where multiple regions contribute collectively to complex cognitive functions. The strong performance of this method suggests that ensemble approaches may be particularly well-suited for capturing the multifaceted nature of brain-behavior relationships in clinical populations.

In contrast, the more limited performance of linear methods such as Ridge regression and linear SVR highlights the constraints of oversimplified approaches to mapping lesions and behaviors. While these methods offer advantages in terms of interpretability and computational efficiency, they appear inadequate for capturing the complex, non-linear relationships that characterize brain-behavior interactions. This finding emphasizes the importance of selecting appropriate analytical approaches that match the complexity of the underlying biological processes being studied. The superior performance of non-linear methods in our study suggests that the relationship between lesion location and aphasia severity involves intricate interactions between brain regions that linear models cannot adequately capture. From a clinical perspective, the identification of critical ROIs through SHAP analysis provides practical guidance for rehabilitation planning and therapeutic intervention. The prominence of the left arcuate fasciculus in our results highlights the importance of white matter integrity in language function, suggesting that rehabilitation strategies should consider not only cortical damage but also the connectivity between language-related brain regions.

The high importance scores for Broca's area underscore the critical role of this region in speech production, suggesting that patients with lesions in this area may benefit from targeted speech therapy interventions that focus on production-based exercises and grammatical processing tasks. The significant SHAP values associated with the right arcuate fasciculus reveal the often-overlooked contribution of right-hemisphere pathways to language function. While traditionally considered to play a more passive or supportive role in language processing, our findings suggest that right-hemisphere structures may be more important for language recovery and compensation than previously recognized. This has important implications for rehabilitation approaches, as therapies that engage both hemispheres might be more effective than those focused solely on left-hemisphere structures. The bilateral nature of language processing, as revealed through our analysis, suggests that rehabilitation protocols should consider the potential for cross-hemispheric compensation and plasticity.

The elevated importance of the left superior temporal gyrus, as revealed through SHAP analysis, emphasizes the critical role of temporal lobe structures in language comprehension. This finding aligns with established knowledge about the superior temporal gyrus's role in auditory processing and phonological analysis, key components of language understanding. Clinically, this suggests that patients with damage to this region may benefit from auditory training and phonological awareness interventions as part of their rehabilitation program. Conversely, the relatively lower importance scores for regions such as the occipital fusiform gyrus reinforce the recognized dominance of frontal and temporal areas in core language processing functions. This finding provides validation for focusing rehabilitation efforts on the most critically affected regions while also highlighting the hierarchical nature of language processing in the brain. The differential importance of various brain regions, as quantified through SHAP values, offers a data-driven approach to prioritizing therapeutic interventions based on the specific lesion patterns of individual patients.

The clinical translation of these findings represents a significant advancement in personalized neurorehabilitation. By providing quantitative measures of each brain region's contribution to language deficits, SHAP-based analysis enables clinicians to develop more targeted and effective rehabilitation strategies. The ability to identify which brain regions contribute most significantly to an individual patient's deficits allows for the customization of therapy protocols based on specific lesion patterns, potentially leading to more efficient and effective rehabilitation outcomes. This personalized approach to neurorehabilitation represents a paradigm shift from one-size-fits-all interventions to precision medicine approaches that

account for individual neuroanatomical variations and patterns of damage. Furthermore, the integration of SHAP analysis with lesion-symptom mapping provides a framework for monitoring treatment progress and adjusting therapeutic interventions based on quantitative measures of brain-behavior relationships. As patients progress through rehabilitation, changes in the relative importance of different brain regions could potentially guide modifications to treatment protocols, ensuring that therapy remains optimally targeted throughout the recovery process. This dynamic approach to rehabilitation planning represents a significant advancement in clinical practice. It holds promise for enhancing outcomes for patients with aphasia and other neurological conditions that affect language function.

6. Conclusion

This research proposes an ROI-based LSM framework augmented with Explainable AI for predicting aphasia severity. From high-resolution MRI data and lesion masks aligned to the MNI template, ROI overlap features were extracted, and seven machine learning models were benchmarked. SVR model with an RBF kernel emerged as the best-performing model, explaining nearly 92.7% of the variance in aphasia scores. SHAP analyses identified Broca's Area and the Left Arcuate Fasciculus as major contributors, delivering insights in a transparent and patient-centered manner. In the broader context of neuroimaging and AI, integrating ROI-based LSM with SHAP bridges the gap between high-performing models. Future studies might expand these findings by incorporating larger patient cohorts, exploring additional ROIs, or integrating automated hyperparameter tuning to refine model performance further. These advancements could inform targeted rehabilitation strategies and enhance clinical decision-making.

Acknowledgement: The authors sincerely thank their colleagues for their valuable support and encouragement. Their guidance and facilities greatly contributed to the successful completion of this work.

Data Availability Statement: This research presents a Shapley value-based framework for transparent machine learning in ROI-based lesion-symptom mapping of aphasia. The data supporting the findings can be made available upon reasonable request to the corresponding authors.

Funding Statement: The authors declare that no funding was received for the conduct of this research.

Conflicts of Interest Statement: The authors declare that they have no conflicts of interest. All citations and references are properly acknowledged based on the information utilized.

Ethics and Consent Statement: Ethical approval was obtained, and informed consent was received from the organization and all individual participants during data collection.

References

1. E. Bates, S. M. Wilson, A. P. Saygin, F. Dick, M. I. Sereno, R. T. Knight, and N. F. Dronkers, "Voxel-based lesion-symptom mapping," *Nature Neuroscience*, vol. 6, no. 5, pp. 448–450, 2003.
2. C. Rorden, H. O. Karnath, and L. Bonilha, "Improving lesion-symptom mapping," *J. Cogn. Neurosci*., vol. 19, no. 7, pp. 1081–1088, 2007.
3. K. Najarian, D. Kahrobaei, E. Dominguez, and R. Soroushmehr, "Artificial Intelligence in Healthcare and Medicine," Boca Raton, *CRC Press*, Florida, United States of America, 2022.
4. S. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," *arXiv preprint*, arXiv preprint arXiv: 1705.07874, 2017, Available: https://arxiv.org/abs/1705.07874. [Accessed by 12/06/2024].
5. F. Nabizadeh and M. H. Aarabi, "Functional and structural lesion network mapping in neurological and psychiatric disorders: a systematic review," *Front. Neurol*., vol. 14, no. 6, pp. 1-25, 2023.
6. G. Varoquaux, "Nilearn: Machine learning for neuroimaging in Python," 2016. Available: https://github.com/andreas-koukorinis/gaelvaroquaux.github.io/blob/master/user_guide.html [Accessed by 12/06/2024].
7. R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. R. Stat. Soc. Ser. B Stat. Methodol*., vol. 58, no. 1, pp. 267–288, 1996.
8. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, no. 10, pp. 2825–2830, 2011.
9. A. Abraham, F. Pedregosa, M. Eickenberg, P. Gervais, A. Mueller, J. Kossaifi, A. Gramfort, B. Thirion, and G. Varoquaux, "Machine learning for neuroimaging with scikit-learn," *Frontiers in Neuroinformatics*, vol. 8, no. 2, pp. 1–10, 2014.

10. A. J. Smola and B. Schölkopf, “A tutorial on support vector regression,” *Stat. Comput*., vol. 14, no. 3, pp. 199–222, 2004.
11. M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mane, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viegas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, “TensorFlow: Large-scale machine learning on heterogeneous distributed systems,” *arXiv preprint* arXiv:1603.04467, 2016. Available: https://arxiv.org/abs/1603.04467. [Accessed by 12/06/2024].
12. L. Breiman, “Random forests,” *Mach. Learn*., vol. 45, no. 1, pp. 5–32, 2001.
13. T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” *in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD ’16)*, San Francisco, CA, United States of America, 2016.
14. J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” *Ann. Stat*., vol. 29, no. 5, pp. 1189–1232, 2001.