Impact of Imputation Methods on Classifying Mouse Primary Visual Cortex Inhibitory Neurons

Author: Orkun Sefik

Vrije Universiteit Amsterdam, Computer Science Department, The Netherlands

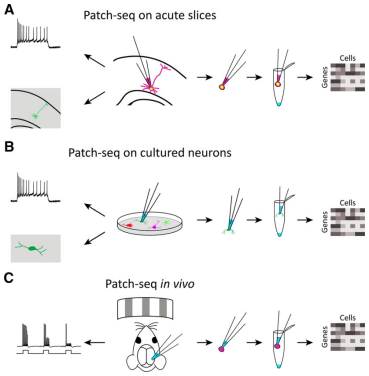
Supervisor: Aneta Lisowska
2nd Assesor: Mark Hoogendoorn

Abstract. In recent years, single-cell multimodal data, such as that obtained through Patch-seq, has enabled increasingly detailed classification of neuron subtypes. However, missing data across transcriptomic, electrophysiological, and morphological features introduces uncertainty into subsequent analyses, including classification. This research investigates how various imputation strategies influence the classification of inhibitory neurons from the mouse primary visual cortex. Focusing on neuron types Lamp5, Pvalb, Sncg, Sst, and Vip, from a widely used dataset of mouse visual cortex neurons, I compare imputation methods ranging from traditional statistical techniques to modern deep learning-based approaches. I hypothesized that deep-learning based imputation methods will generally outperform traditional techniques in preserving biologically meaningful variability and subtype distinctions. Specifically, I hypothesized that accurate imputation improves classification performance and reduces uncertainty in subtype assignment. By systematically evaluating these methods, this work aims to inform best practices in preprocessing noisy multimodal neuroscience data and enhance the biological interpretability of resulting classifications.

Keywords: machine learning · neuron classification · imputation methods · Patch-seq · missing data · multimodal analysis · computational neuroscience · biological interpretability

1 Introduction

Advances in single-cell analysis have developed neuroscience by enabling multimodal profiling of neurons, uncovering the diversity of cell types in the brain. Techniques like Patch-seq allow researchers to simultaneously capture transcriptomic, electrophysiological, and morphological data from single neurons, providing a new and comprehensive view of how cellular identity is characterized. [5]. One of the most significant advancements resulting from this progress has been the ability to perform multimodal profiling where multiple types of biological data are collected from the same single cell. This multimodal approach is particularly valuable in neuroscience, where understanding a neuron's identity often requires integrating information across different dimensions: transcriptomic (gene expression), electrophysiological (firing patterns), and morphological (shape, structure and connectivity). Among these, the inhibitory neurons within the mouse primary visual cortex have been extensively investigated due to their significant contribution to the regulation of cortical activity. Datasets such as Gouwens et al. [1] provide rich multimodal information, but are often negatively affected by missing values, a common issue in high-dimensional and multimodal biological datasets like Patch-seq [6]. These missing values may arise from technical failures during cell capture, limitations in sequencing depth, or loss of information during morphological mapping.



The use of machine learning (ML) to classify neurons into known cell types offers great potential for neuroscience research, particularly when complex features captured by Patch-seq are being used. ML is already demonstrating practical impact in medical imaging and diagnostics, for example, a patented method based on symmetry-aware deep learning has been developed to detect early signs of stroke in brain CT scans [US10163040B2], showing how computational models are being applied to extract subtle patterns in biological data [29]. Still, the reliability of these models heavily depends on how missing data is addressed, since it is well-known that the presence of missing values is a negative effect on classification tasks [33]. Imputation, the process of estimating and filling missing values, is often seen as the standard preprocessing step. However, despite evidence that imputation strategies can significantly affect the accuracy of cell type identification [7], its influence on subsequent classification outcomes is rarely analyzed. This project investigates how different imputation techniques impact neuron classification, a question of both computational and biological relevance. The goal is not only to compare imputation methods but to understand their implications for the scientific interpretation of neuronal identity.

2 Background

Cell types are the basic functional units of an organism. Cell types display diverse phenotypic properties at multiple levels, making them challenging to define, categorize, and understand [22]. Figuring out what makes one type different from another helps us understand how the brain works at a fundamental level. In this project, the focus is on inhibitory neurons from the mouse visual cortex, specifically the following subclasses: Lamp5, Pvalb, Sncg, Sst, and Vip. These subclasses represent distinct inhibitory interneurons, each defined by different characteristics, molecular profiles and electrophysiological properties.

The dataset is rich but also noisy, with missing values in different parts. That's a problem since missing data can alter the classification accuracy significantly [2]. This issue is particularly critical in multimodal studies, where the absence of one data type (e.g., gene expression) for a subset of cells can distort comparisons and reduce the reliability of machine learning models trained to classify or predict cell types [3]. To deal with that, imputation is widely used by researchers [4]. That's why, despite the remarkable potential of multimodal single-cell datasets like those from Patch-seq, their full value can only be realized when we apply suitable computational strategies to handle the missing data that almost always comes with them. This observation leads to the research question of this study:

"How do different imputation methods affect the classification of these neuron types using machine learning?"

3 Literature Review & Related Work

3.1 AI/ML Methods and Imputation

In the field of single-cell data analysis, machine learning has become a crucial approach for uncovering patterns within high-dimensional, noise-prone datasets. To give an example, machine learning has been successfully applied in identifying cancer cell subpopulations, analyzing therapy-resistant cells, and uncovering gene expressions in tumor microenvironments through single-cell RNA sequencing (scRNA-seq) data [8]. However, a significant challenge in these datasets, especially in multimodal contexts like Patch-seq, is the missing data [6]. Numerous imputation techniques have been proposed to address this issue in both single-cell and biomedical datasets. Classical methods, such as mean or median imputation, are simple and widely used but often fail to account for the underlying data structure. For example, basic tasks like KNN-impute estimates missing values based on the similarity between neighboring samples, also being used in biological contexts [9]. To go beyond these baseline approaches, more recent deep learning-based methods have been developed:

- **DCA,** a denoising autoencoder optimized for scRNA-seq data. [10]
- **MICE,** a statistical imputation method that models each variable with missing values as a function of other variables in an iterative chained sequence. [11]

- **GAIN,** a GAN-based imputation strategy that learns to recover missing values through adversarial learning. [12]
- **CL-Impute,** which uses contrastive learning to guide biologically meaningful imputation. [13]
- **SoftImpute,** a matrix factorization-based technique that estimates missing entries by iteratively approximating the data matrix with a low-rank decomposition, making it particularly suitable for sparse, high-dimensional datasets like those found in single-cell experiments. [14]

Despite these advancements, a critical gap remains in understanding how these imputation choices affect key following tasks such as cell type classification. Studies, including that of Ly et al. [15], have shown that imputation can significantly alter inferred gene-gene relationships, thereby also altering subsequent biological interpretations. Yet, in most studies, imputation methods are chosen based on computational performance metrics rather than their biological impact on tasks such as classification. The term 'biological impact' refers to the influence of imputed data on the interpretation of underlying biological patterns. In many studies, imputation methods are selected based on how fast or accurate they are computationally, but less attention is given to whether the results remain biologically meaningful or lead to correct conclusions in tasks like classifying cells or understanding disease mechanisms. This is especially relevant for studies that aim to link transcriptomic data with other modalities, such as morphology and electrophysiology, where each modality might suffer from its own type and pattern of missingness.

3.2 Neuroscience: Data, Features, and Biological Context

In neuroscience, classifying neurons into cell types is essential for making sense of the brain's functional architecture. A powerful technique that supports this classification is Patch-seq, allowing for rich multimodal characterization of cell types. Originally developed for use in mouse brain slices, it has become foundational in neuroscience studies that aim to connect gene expression with functional properties at the cellular level [5]. Gouwens et al. [1] used this technique to profile inhibitory interneurons in the mouse visual cortex and identified well-defined transcriptomic subclasses such as Lamp5, Pvalb, Sncg, Sst, and Vip. Each of these groups shows characteristic patterns across features, showing how each type of cell has unique variations in features. In Patch-seq and other multimodal contexts, datasets are often affected by missing values in one or more modalities, which introduces challenges for analyses such as classification. Scala et al. [16] further investigated the multimodal structure of the mouse motor cortex and observed that, although broad cell families such as Pvalb and Sst remain clearly distinguishable, there exists a continuous gradient of variability within each individual group. This suggests that classification may not always be evident, especially when features like morphology or electrophysiology are incomplete. Lee & Dalley et al. [17] extended this work to human GABAergic neurons, showing both conserved and unique properties compared to mice. Their

study highlighted how missing data and slight subclass differences complicate the task of building reliable classifiers. In the context of disease, the implications become even more significant. The SEA-AD study [18] found that in Alzheimer's disease, certain inhibitory neuron subclasses, specifically Vip+ and Pvalb+, are selectively vulnerable. Accurate classification is crucial not just for taxonomy but especially for identifying which cell types are most affected in neurodegenerative processes. In this context, accurate classification of neuron types is more than just an academic exercise, it has real clinical implications. Classification in medical sciences is very vital as it is a matter of life or death [26]. Misclassifying vulnerable subtypes could mean failing to detect some mechanisms of cognitive decline or potential targets for therapy in neurodegenerative diseases like Alzheimer's.

3.3 Combined Approaches and Gaps in the Literature

Bridging neuroscience and AI/ML is a still developing area. While machine learning models have been used to classify neurons based on electrophysiology or transcriptomics, few studies have explored how imputation affects these classifications. One study done by Vasques et al. [19] focused on morphological neuron classification using machine learning. They assessed various algorithms to classify neurons based on morphological features extracted from histological reconstructions. Their findings showed that supervised methods, particularly linear discriminant analysis, achieved superior classification performance, highlighting the importance of quality data and algorithm selection. For example, Scala et al. [16] and Lee & Dalley et al. [17] also used high-quality Patch-seq data but didn't examine how imputation might bias cell type assignments. Scala et al. [16] used k-NN as their main imputation method to keep the data directly related to their t-SNE embeddings. Lee & Dalley et al. [17] also used k-NN to impute their data. One study showed this potential by developing a domain-adaptive neural network, which is capable of classifying neurons across species, using only electrophysiological data [20]. Still, the impact of imputation choices on classification remains an open and important question, especially in complex, multimodal datasets like Patch-seq.

Moreover, imputation strategies are often applied without considering the biological structure of the data. In multi-modal settings, this can lead to inconsistencies between what is reconstructed and what is biologically plausible. This is especially problematic when missing data isn't random, for instance, when morphological reconstructions fail more often in certain cell types. Despite the importance of this issue, there's a clear lack of studies that compare different imputation methods in a controlled, biologically grounded setting. While many studies implement imputation without systematically evaluating its influence on subsequent analyses, an increasing number have started to acknowledge the potential implications of the preprocessing choice. For instance, studies by Shadbahr et al. [21] and Jäger et al. [35] discusses the importance of robust imputation methods in the context of machine learning, emphasizing that the choice of imputation technique can significantly influence model performance. Despite

these novelties, there still remains a lack of systematic evaluations comparing different imputation methods in the context of neuronal classification tasks. This study aims to fill this gap by evaluating multiple imputation techniques on inhibitory neurons from the mouse primary visual cortex, assessing how these choices impact the classification performance both computationally and biologically. Rather than simply identifying the most accurate technique, the objective is to understand how different imputation choices shape the resulting biological interpretations. This gap suggests the need for a systematic comparison of imputation methods in a biologically grounded setting. This study addresses this by evaluating classical and modern imputation strategies for their impact on the classification. By doing so, I aim to guide future neuroscience pipelines toward more reliable and biologically meaningful preprocessing strategies.

4 Methods

4.1 Data and Preprocessing

The dataset includes inhibitory neurons from the mouse primary visual cortex and includes transcriptomic, electrophysiological, and morphological features, originally containing 4435 samples, including both inhibitory and excitatory neurons. However, excitatory neurons were excluded during preprocessing, following the recommendation of an expert neuroscientist supervising this research. This decision was based on two main considerations. First, excitatory neurons were severely underrepresented compared to inhibitory types, which could introduce class imbalance and alter model training. Second, excitatory neurons display some differences compared to inhibitory neurons, potentially affecting model parameters and weights. Despite the availability of various missing value imputation methods, the presence of outliers reduces both the precision of the imputation process and the reliability of biomarker identification [34]. Since the aim of this study is to focus on capturing the very subtle differences among inhibitory neuron subtypes, excitatory neurons were excluded. Lastly, rows with $\geq 50\%$ missing values were dropped. After filtering out the excitatory cells, the final dataset contains 4243 inhibitory neurons classified into five subclasses: Sst (1913), Vip (1026), Pvalb (771), Lamp5 (472), and SNCG (61), see Appendix A for subclass counts and percentages. This led to the whole dataset containing 8.59% missing values.

As a preliminary step, all random seeds and stochastic parameters were set to a fixed value of 31, which was also chosen randomly, in order to eliminate the influence of randomness on the experimental outcomes [23]. Following that the dataset was split into 2 (90% - 10%, train/validation and test set). The test set contains 426 samples, exactly 10% of each subclass and also the missing value percentage is also kept (8.59%). A representative imputation strategy was required to develop the model without introducing bias from relying on a single method. To achieve this, a combination of imputation techniques was used based on missing value thresholds: features with less than 5% missing values were imputed with the mean because the missing portion is small enough that the mean

won't significantly distort the feature's overall distribution, those with 5%–20% missing were imputed with the median due to greater sensitivity to outliers, and features with more than 20% missing were dropped to avoid introducing bias or compromising data integrity. This method includes a combination of different imputation methods, supporting a more generalizable model. After that, highly correlated feature pairs ($|\text{correlation}| \geq 0.9$, see Appendix A for the correlation matrix) are removed by keeping the more important one based on feature ranking. It was determined as shown in Figure 1:

Feature 1	**Feature 2**	**Correlation**
FAP_rheobase	TS1_rheobase	0.986104
FAP_rheobase	TS2_rheobase	0.958476
TS1_rheobase	TS2_rheobase	0.977700

Fig. 1. Highly Correlated Feature Group

In this case, FAP_rheobase, TS1_rheobase, and TS2_rheobase are all highly correlated. So, with respect to our rule, the most important feature would be kept. As seen in Figure 2, FAP_rheboase has a higher importance than the other features, so that one is kept, and the other two (TS1_rheobase and TS2_rheobase are dropped).

Feature	**Importance**
FAP_rheobase	0.030343
TS1_rheobase	0.021526
TS2_rheobase	0.016800

Fig. 2. Importances of a Highly Correlated Feature Group

Dimensionality reduction is analyzed via PCA and UMAP (see Appendix A), followed by an evaluation of the results. These evaluations helped visualize the multi-dimensional dataset in a simpler form, making it easier to observe potential groupings or patterns among neuron types.

4.2 Model Development and Selection

A variety of models were evaluated using AutoGluon with 5-fold stratified cross-validation and shuffled inputs [24]. The best-performing models were Neural

Network, XGBoost, and CatBoost. Ultimately, XGBoost was selected due to its superior consistency and interpretability [25]. The model used is XGBClassifier, from the XGBoost library. Hyperparameter tuning included exploration of different values of "k" (number of neighbours in SMOTE), learning rates, and early stopping rounds. The model was trained on the training set (90%, which also includes the validation set) with the separate test set (10%) to ensure an unbiased final evaluation. SMOTE was used during training folds (excluding the majority class) to address class imbalance, explicitly avoiding its application to the test set to prevent data leakage [30].

4.3 Assessing the Robustness of the Model

The cross-validation resulted in individual fold accuracies between 91.12%, and 93.08%, with a mean of 92.06% and a standard deviation of 0.0077, showing strong stability. Macro F1 average is also computed as 84%. The classification report for the full training data showed high precision and recall ($> 86\%$) for major classes (Lamp5, Sst, Vip, Pvalb), while minority class performance (SNCG) remained modest ($\sim 50\%$), which is expected due to its limited representation. Confusion matrices were analyzed to capture class-level misclassifications, see Appendix B. The test set evaluation further confirmed model performance, with a high accuracy of 92%, and a strong macro F1 score of 0.86. The test set outperforming the training set in this metric can be seen as a positive sign of good generalization, suggesting that the model is not overfitting and is capable of capturing broader patterns that extend beyond the training data. This pipeline is carefully designed to keep the test results honest and reliable. By applying SMOTE only to the training data within each fold, and never to the validation or test sets, it avoids any artificial inflation of performance (eg. data leakage [30]). This separation ensures that what the model sees during training doesn't leak into testing, preserving the integrity of the evaluation and making the results truly reflective of how the model would perform in real-world scenarios.

4.4 Selection of Imputation Methods

To systematically evaluate the effect of imputation methods, the following methods were applied to the original dataset:

- **Mean-only imputation:** Missing values are replaced with the mean of the values for each feature.
- **Median-only imputation:** Missing values are replaced with the median of the values for each feature.
- **k-NN imputation:** Missing values are imputed based on the values of the k (n_neighbors = 5) most similar data points.
- **Constant outlier value imputation:** A fixed value well outside the normal range is used to fill missing entries.
- **Random non-outlier value imputation:** This approach randomly selects a valid, non-outlier value by choosing an already existing value from the same feature to replace missing values.

- **Rule-based hybrid imputation:** This approach applies mean or median imputation based on feature-level missingness thresholds and drops features with excessive missing data.
- **MICE (Multiple Imputation by Chained Equations) [11]:** MICE imputes missing values by iteratively modeling each incomplete feature as a function of the other variables, producing statistically decent estimates based on the assumption that the data are missing at random.
- **Soft Impute [14]:** This method estimates missing values by reconstructing the data matrix using low-rank approximation, accurately identifying some patterns within the data.

Each imputed dataset was passed through the same trained XGBoost model, label encoder and scaler, ensuring all differences in performance can be linked only to the imputation method, and not to other variables. Also the imputer and the input columns for each method were stored, and then imported again for the testing to keep consistency.

4.5 Experimental Pipeline

The dataset was splitted into 5 pairs, each representing a different 80-20 split, thus covering the whole data. The train and test sets were paired in a way that they also always contained respectively 80% and 20% of each subtype. Also, before starting the experiments, the amount of missing values was increased (randomly chosen cells, excluding subtypes and cell ID) in the training sets. This resulted in the training sets containing $\sim$ 24.3% missing values, while the test sets contain $\sim$ 8.59%. This was done to increase the effect of the imputation methods on the evaluation metrics in the future. The process begins by encoding the target variable subclass using LabelEncoder, which encodes the categorical class labels into numerical labels (in this case: 1-5), suitable for model training. Features such as the raw subclass label and cell IDs are excluded to prevent leakage. A stratified 5-fold cross-validation strategy is applied to ensure each fold preserves the original class distribution, which is especially critical given the class imbalance in the dataset. Within each fold, the data is first split into training and validation subsets, followed by feature scaling using StandardScaler to normalize the input range and facilitate faster model convergence. The data is standardized to have zero mean and a standard deviation (std) of 1. To address class imbalance during training, SMOTE is applied with the condition sampling strategy "not majority", meaning all classes except the majority one are oversampled [28]. This process minimizes class imbalance in the training data, while keeping the feature space fundamentally the same. Importantly, the imputation parameters were learned only from the training set and then applied to the test set without recalculating them. This approach prevented any possible data leakage. After training, the model makes predictions on the validation set within each fold. True and predicted labels are taken from all folds to calculate overall performance metrics. Accuracy for each fold is printed to keep track of how stable the model is across different subsets of the data, followed by the mean accuracy and standard

deviation. To have a clearer picture of how the model performs across each class, a detailed classification report and confusion matrix are also generated, providing key aspects of class-wise performance. The training set was imputed using the respective method. For illustration, consider the mean-only imputation approach:

Each paired set creates a model. So; the model used with the train1 - test1 pair, is called meanmodel1 in this case. The training set was first imputed using fit_transform, as the imputer needed to be fitted on the training data. For example, if a column contains values such as [1, 3, 5, 7, NaN], the missing value would be filled in with 4, the mean. Following imputation, the XGBoost model was trained using hyperparameters identified in the preprocessing phase. As explained above, for the test set, the previously fitted imputer was used via its transform method only, not re-fitted. This design choice avoids any form of data leakage. In the current example, even if the test set has a column like [15, 16, NaN, 12, 14], the missing value would still be filled with 4, not recalculated. Right after those, the imputed test set is given to the trained model, calling the same scaler, and label encoder as well. All 5-folds are used to predict by their respective model. Following this, the average and standard deviations of the evaluation metrics for each method are calculated.

5 Results

To develop an initial baseline model and establish a general workflow, the dataset was first split into 90% training and 10% testing, as described earlier in Section 4.1. This step enabled model selection and hyperparameter tuning under standard conditions. After finalizing the model architecture, a separate evaluation of imputation methods was conducted. For this purpose, the data was re-split into 80% training and 20% testing, and a 5-fold cross-validation (CV) was performed on the training portion, as outlined in Section 4.5.

Model performance was assessed using standard metrics (accuracy, precision, recall, macro F1, and class-wise F1 scores). In addition, statistical significance between methods was tested using p-values. These p-values then were compared against every other imputation method used, to show any significant difference available. To interpret how imputation affects model behavior in detail, and to get some biological insight, SHAP values were computed [27].

The highest classification accuracy was observed with k-NN (92.03%) and soft imputation (92.01%), followed by mean, median, and MICE. The constant outlier method consistently performed worst (89.18%), indicating it is, expectably, not a good option for biological datasets where preserving structure is important. Precision was highest for random imputation method, this may be caused by models trained on datasets imputed more randomly may be more prone to predicting different subclasses instead of relying on the majority classes. Recall was highest for k-NN and MICE. Macro-F1 scores showed a similar trend: k-NN (81.65%) and MICE (81.25%) led, with all others ranging from 78.59% to 80.26%, while the constant method dropped significantly to 74.36%. Together,

these two metrics indicate that k-NN and MICE provide the most stable and balanced classification performance across major and minority subclasses.

5.1 Statistical Comparison of Accuracy

Pairwise p-value analysis confirms that the constant outlier method performs significantly worse than all other imputers (p ¡ 0.0001 across the board). This shows the danger of using biologically implausible values to fill in missing data. Among better-performing methods, k-NN significantly outperforms both random ($p = 0.0043$) and rule-based ($p = 0.0047$) imputers, while soft imputation also shows a significant edge over those two ($p = 0.0409$ and $p = 0.0268$, respectively). No statistically significant accuracy differences were observed between mean, median, and MICE, suggesting these approaches are largely interchangeable in terms of raw accuracy. However, the clear advantage of distance-based and matrix factorization approaches like k-NN and soft imputation highlights their robustness. (figure 3)

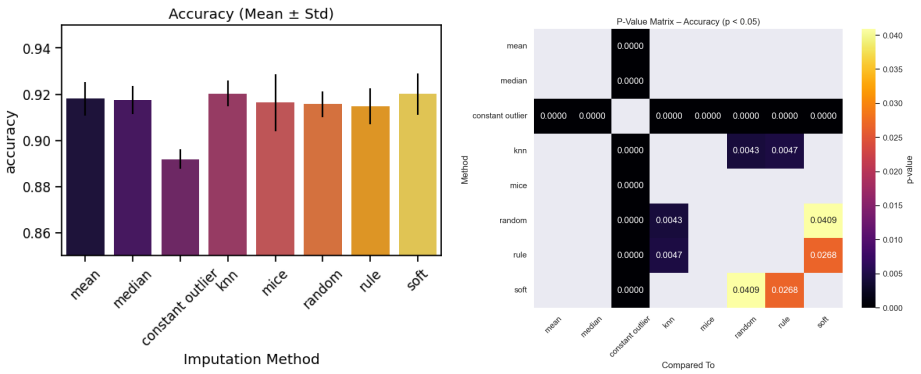


Fig. 3. Comparison of classification accuracy across different imputation methods (left) and pairwise statistical significance of differences using p-values (right)

5.2 Precision and Recall

Precision and recall generally reinforce the conclusions drawn from accuracy. The constant outlier method remains significantly weaker in both metrics. Precision was highest for the random method, potentially due to increased diversity in feature values, whereas recall was highest for k-NN and MICE. Some significant pairwise differences appeared, particularly involving soft and rule-based imputers, but overall, these metrics had a limited effect on method rankings. To summarize this, in terms of precision, these imputers may perform similarly enough to be considered practically equivalent; while recall keeps the constant method

as a weak outlier and shows some new significant differences among common methods, highlighting the importance of careful imputer selection when recall is critical, such as in medical or fraud detection tasks [31] [32].

5.3 Statistical Comparison of Macro-F1

Macro-F1 analysis again places the constant outlier method significantly below all others. The highest scores were from k-NN and MICE, with no significant difference between them ($p = 0.4721$), confirming their ability to handle class imbalance effectively. Soft imputation, although strong, shows statistically significant differences from both k-NN ($p = 0.0126$) and MICE ($p = 0.0295$), suggesting its balanced performance doesn't match that of the top two. Even among simpler methods, meaningful differences exist: for example, mean vs. k-NN ($p = 0.0294$) and rule vs. mean ($p = 0.0755$) show that not all basic imputers behave similarly. Macro-F1 therefore captures small performance differences that overall accuracy may hide, which is crucial in biological contexts where investigating differences between close subclasses is essential. (figure 4)

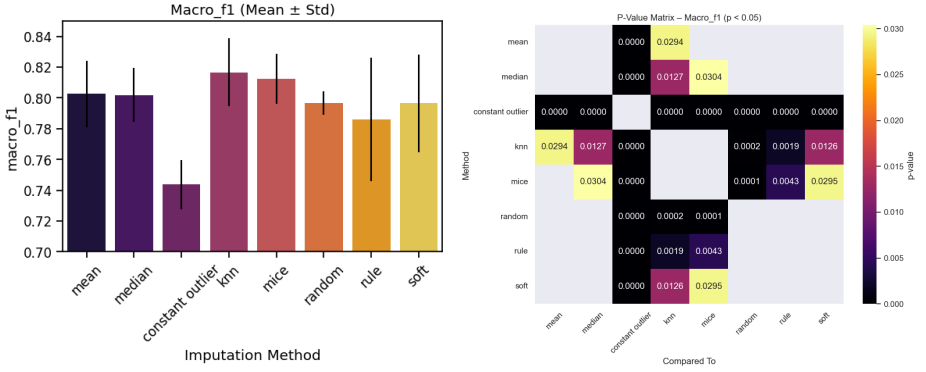


Fig. 4. Comparison of Macro-F1 across different imputation methods (left) and pairwise statistical significance of differences using p-values (right)

5.4 Individual Subclass Performances

Although the main performance metric reported in this study is the macro-averaged F1-score, the model's performance across individual subclasses was also analyzed. However, due to the significant class imbalance (SNCG cells constitute less than 2%), reporting only class-wise accuracies could be misleading. Therefore, macro-F1, which equally weights all classes, was chosen as the main evaluation metric. Most of the models showed high macro-f1 scores ($\geq 89\%$) in all subclasses except SNCG, which was the extreme minority class. Even though

it's not the majority class, Lamp5 had an average F1 of 93.5%, leading all other subclasses. Pvalb returned an average F1 of 89.5%. Sst, the majority class, had an average F1 of 93%. Vip showed an average F1 of 91.63%. However, performance on SNCG (n = 61) was lower compared to the other subclasses, regardless of imputation strategy, with an average macro F1-score of 30.75%. Even though a carefully adjusted SMOTE used to help balance the training data, the test sets were designed by keeping the natural class proportions, causing even a couple labelling errors leading to this drop in minority class.

5.5 SHAP Analysis

SHAP values were used to analyze feature importances for each cell class across imputers and to better understand how different imputations affect model interpretation. Results showed that feature influence varies by class and imputation method, indicating that data completion can subtly affect model predictions. To illustrate, SHAP plots for Class 0 (Lamp5) and Class 2 (SNCG) using k-NN and mean imputation from fold 3 are shown below. These two classes were selected to highlight contrasting scenarios: Lamp5 had the highest prediction performance, while SNCG had the lowest. The remaining classes are discussed briefly in the text to maintain conciseness and avoid redundancy. Fold 3 was selected arbitrarily among the five cross-validation folds, as there is no systematic pattern or correlation between fold number and model performance.

For Lamp5 (Class 0), features such as FAP_halfwidth, tau and TS2_halfwidth, were most impactful under k-NN imputation. Higher FAP_halfwidth and lower input tau increased the model's confidence in predicting Lamp5, while average values of TS2_haldwidth slightly pushed the model to predict Lamp5 [figure 5]. Under mean imputation, similar features remained important with similar influences on model prediction, but SHAP values were more tightly clustered, suggesting reduced variability and sharper decision boundaries [figure 5].

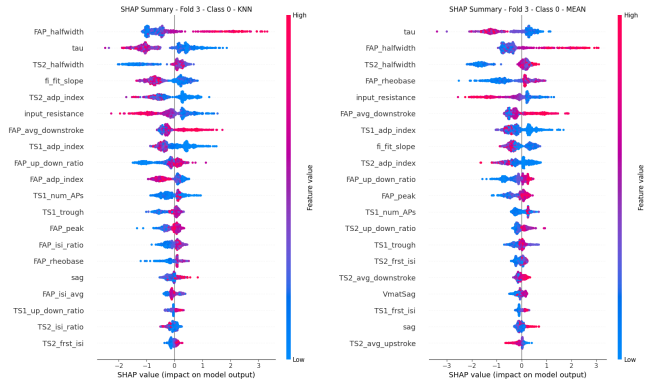


Fig. 5. SHAP Analysis of Lamp5, using KNN (left) and mean imputation (right)

In Vip (Class 1), features like FAP_up_down_ratio, FAP_peak, and TS1_isi_avg played the largest roles in k-NN imputation. High values of these features generally pushed predictions away from this class. Mean imputation produced a smoother SHAP distribution, varying some features (TS1_up_down_ratio and FAP_halfwidth rose), with the same influences. For SNCG (Class 2), the three most influential features under k-NN were FAP_up_down_ratio, TS1_isi_avg, and tau, all showing directionally consistent patterns in reducing or increasing the likelihood of a SNCG prediction [figure 6]. Mean imputation raised the role of TS1_up_down_ratio, TS2_adp_index and vmbaseM, showing that the model's feature importance changed depending on the imputation method [figure 6].

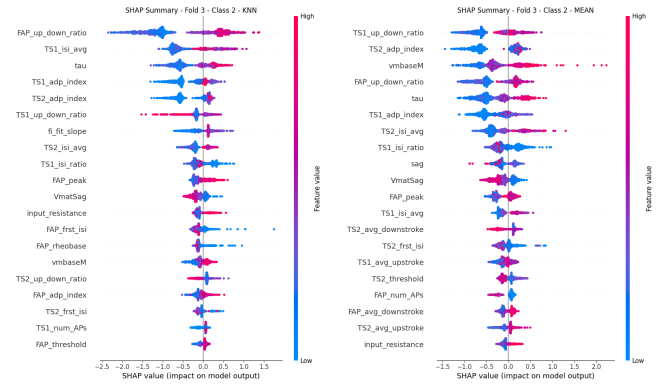


Fig. 6. SHAP Analysis of SNCG, using KNN (left) and mean imputation (right)

In Sst (Class 3), FAP_up_down_ratio, tau and TS1_up_down_ratio were again dominant, emerging as consistent top predictors across both imputers. Finally, for Pvalb (Class 4), the features TS2_up_down_ratio, TS1_trough, and also the feature TS1_up_down_ratio was the top features with k-NN imputer, with the higher the value the more the model predicted Pvalb. However, mean imputation broadened the feature influence, with FAP_rheobase gaining importance.

6 Discussion

The results section conveys a clear message: imputation strategy is not a simple preprocessing detail but a main determinant of subsequent model behaviour. The consistent poor performance of the constant outlier method shows that using biologically meaningless values to impute the data alters the natural relationships between data points, a deformation the classifier cannot fully unlearn even with generous data augmentation. In contrast, distance-based (k-NN) and low-rank-based (soft) imputers kept the natural groupings in the data as well as hidden feature relationships delivering both the highest accuracies and the

most interpretable SHAP-based feature profiles. That k-NN and MICE achieved the strongest recall, while random imputation gave a slight increase to precision, is a good reminder that model performance is rarely one-dimensional. What counts as superior depends heavily on context. In scenarios where missing data aren't random, and where rare cell types could be crucial to understanding disease, recall matters more than precision [31] [32]. Macro-F1 highlighted the strength of k-NN and MICE, but the significant advantage of soft imputation shows that matrix factorization deserves more attention in electrophysiological pipelines. Subclass-level results further deepened these insights. High F1 values for Lamp5, Vip, Sst, and Pvalb indicate that, once sufficient data is available, most reliable imputation methods generally lead to consistent model decisions Yet the consistently under-represented SNCG class revealed the vulnerabilities of all methods tested, including the most reliable method: k-NN. Even after applying SMOTE, the sharp decline in SNCG's F1 score shows a basic limitation: no imputation technique can create useful information out of near-empty data. The interpretability results showed a more detailed insight. Some electrophysiological features stayed consistently influential across imputers, but their relative importance shifted in small, and sometimes unintuitive ways. However, those shifts matter: a researcher inferring biological mechanisms from SHAP rankings might land at different biological interpretations depending on the chosen imputer. For this reason, I argue that any study including feature importance must explicitly disclose its imputation strategy, ideally supported by sensitivity analyses that show how interpretations could shift. Two other limitations deserve acknowledgement. First, while five-fold cross-validation gives a reasonable estimation of how well the models might generalize, relying on just one dataset limits ecological validity. Electrophysiological patterns can vary a lot depending on brain region, developmental stage, and other biological factors. Second, I treated imputers as just simple tools, intentionally keeping hyperparameter tuning minimal to keep variables consistent and avoid introducing unintended bias. A more detailed grid search could reveal even larger variation in results or reduce the performance differences. Future work could explore targeted oversampling or generative augmentation specific to SNCG-like minorities, potentially guided by SHAP values. Future work could also include statistical testing for significance of SHAP-derived feature interpretations, particularly with different imputation methods, to better understand the effects of methods on biological meaning. Additionally, a broader range of imputation methods could be evaluated, potentially identifying optimal parameter thresholds or regions that consistently deliver both strong performance and interpretability.

7 Conclusion

This thesis establishes that the choice of imputation method has a measurable, and at times decisive, impact on classifier accuracy, robustness to class imbalance, and biological interpretability. Distance-based (k-NN) and model-based (MICE) imputation methods appear as the most reliable options, with soft im-

putation (SoftImpute) offering an alternative to them. Simpler methods (mean, median, random, rule-based) may be sufficient when speed is prioritized over slight performance improvements, however the constant outlier method should always be avoided. This shows that in multi-modal datasets, simply imputing missing values with outliers and disregarding them is poor practice. This study also highlights that preprocessing decisions play a critical role in biological data science. They alter not only performance but also the biological interpretability of the data. Notably, k-NN imputation preserved biologically meaningful distinctions between cell types by maintaining sharper feature boundaries across classes, thus preserving interpretability and supporting more accurate biological inferences. In contrast, simpler methods like mean imputation tended to smooth out these distinctions, risking the loss of important patterns and limiting the chance to make accurate biological interpretations. This also highlights that the imputation method actively influences how each neuron is represented and structured before being given to the model. If this representation is skewed, the model will learn from data that no longer accurately reflects real-world biological variation. This misrepresentation can lead to misleading conclusions, reduced generalizability, and ultimately, a failure to capture the true complexity of neuronal behavior. Therefore, thoughtful imputation is not just a preprocessing step, but a foundational part of preserving biological meaning in computational neuroscience. This becomes even more critical given the high likelihood that such models will be used in medical applications in the future, where decisions based on biologically inaccurate data could have serious consequences [36]. As datasets become increasingly large and complex, the impact of data preprocessing decisions will only become more significant [37]. Therefore, using thoughtful, evidence-based imputation is a crucial step toward making sure our scientific findings in neuroscience are robust and repeatable. In sum, by systematically testing how imputation methods influence both predictive and explanatory aspects of neuronal classification, this thesis provides a solid foundation for scientists and contributes to a more transparent and methodologically well-informed phase of computational neuroscience. It concludes that imputation, class imbalance, and interpretability should not be treated as completely separate tasks, but rather as interconnected issues that need to be handled together.

8 Acknowledgements

This thesis project was conducted as the Bachelor Project of the BSc Artificial Intelligence program at VU Amsterdam, and also as part of an internship at the Center for Neurogenomics and Cognitive Research (CNCR). I would like to sincerely thank my thesis supervisor, Dr. Aneta Lisowska, my internship supervisor, Dr. Christiaan de Kock, and Femke Waleboer for their unwavering support, generous guidance, and inspiring mentorship throughout the project.

The source code developed for this thesis is hosted on a private GitHub repository and can be accessed upon request.

References

1. Gouwens, N. W., Sorensen, S. A., Baftizadeh, F., et al. (2020). Integrated morphoelectric and transcriptomic classification of cortical GABAergic cells. Cell, 183, 935–953. https://doi.org/10.1016/j.cell.2020.09.057
2. K. Mehrabani-Zeinabad, M. Doostfatemeh, and S. M. T. Ayatollahi, "An efficient and effective model to handle missing data in classification," BioMed Research International, Nov. 25, 2020. https://doi.org/10.1155/2020/8810143
3. M. Amodio, S. E. Youlten, A. Venkat, B. P. San Juan, C. L. Chaffer, and S. Krishnaswamy, "Single-cell multi-modal GAN reveals spatial patterns in single-cell data from triple-negative breast cancer," Patterns, vol. 3, no. 9, p. 100577, Sep. 2022. https://doi.org/10.1016/j.patter.2022.100577
4. J. Xue, "Review on data imputation methods in machine learning," *Journal of Physics: Conference Series*, vol. 2646, no. 1, p. 012034, Dec. 2023. https://scispace.com/papers/review-on-data-imputation-methods-in-machine-learning-38r9aux1f2
5. Cadwell, C., Palasantza, A., Jiang, X. et al. Electrophysiological, transcriptomic and morphologic profiling of single neurons using Patch-seq. Nat Biotechnol 34, 199–203 (2016). https://doi.org/10.1038/nbt.3445
6. Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D. J., Hicks, S. C., Robinson, M. D., Vallejos, C. A., Campbell, K. R., Beerenwinkel, N., Mahfouz, A., et al. (2020). Eleven grand challenges in single-cell data science. Genome Biology, 21(1), 31. https://doi.org/10.1186/s13059-020-1926-6
7. Zand, M., & Ruan, J. (2020). Network-based single-cell RNA-seq data imputation enhances cell type identification. Genes, 11(4), 377. https://doi.org/10.3390/genes11040377
8. K. Asada *et al.*, "Single-Cell Analysis Using Machine Learning Techniques and Its Application to Medical Research," *Biomedicines*, vol. 9, no. 11, p. 1513, Oct. 2021. https://scispace.com/papers/single-cell-analysis-using-machine-learning-techniques-and-55oq9v4ues
9. Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., & Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. Bioinformatics, 17(6), 520–525. https://doi.org/10.1093/bioinformatics/17.6.520
10. Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S., & Theis, F. J. (2019). Single-cell RNA-seq denoising using a deep count autoencoder. Nature Communications, 10, Article 390. https://doi.org/10.1038/s41467-018-07931-2
11. M. J. Azur, E. A. Stuart, C. Frangakis, & P. J. Leaf, "Multiple imputation by chained equations: what is it and how does it work?," *International Journal of Methods in Psychiatric Research*, vol. 20, no. 1, pp. 40–49, 2011. https://doi.org/10.1002/mpr.329
12. J. Yoon, J. Jordon, & M. van der Schaar, "GAIN: Missing Data Imputation using Generative Adversarial Nets," in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018. https://doi.org/10.48550/arXiv.1806.02920
13. Xu, Y., Zhang, H., Ren, X., et al. (2023). CL-Impute: Contrastive learning-based imputation for dropout single-cell RNA-seq data. Computers in Biology and Medicine, 169, 107618. https://doi.org/10.1016/j.compbiomed.2023.107263
14. R. Mazumder, T. Hastie, and R. Tibshirani, "Spectral regularization algorithms for learning large incomplete matrices," *Journal of Machine Learning Research*, vol. 11, pp. 2287–2322, 2010. https://dl.acm.org/doi/10.5555/1756006.1859931

15. Ly, L. H., & Vingron, M. (2022). Effect of imputation on gene network reconstruction from single-cell RNA-seq data. Patterns, 3(2). https://doi.org/10.1016/j.patter.2021.100414
16. Scala, F., Kobak, D., Bernabucci, M., et al. (2020). Phenotypic variation of transcriptomic cell types in mouse motor cortex. Nature, 598, 144–150. https://doi.org/https://doi.org/10.1038/s41586-020-2907-3
17. Lee, B. R., Dalley, R., Miller, J. A., et al. (2023). Signature morphoelectric properties of diverse GABAergic interneurons in the human neocortex. Science, 382, eadf6484. https://doi.org/10.1126/science.adf6484 https://doi.org/10.1126/science.adf6484
18. Seattle Alzheimer's Disease Cell Atlas (SEA-AD). (2024). Integrated multimodal cell atlas of Alzheimer's disease. Nature Neuroscience, 27, 2366–2383. https://doi.org/10.1038/s41593-024-01774-5
19. Vasques, R., Esteves, P., Ribeiro, A., & Caetano, T. (2025). Neuronal morphology classification: A machine learning approach. arXiv. https://arxiv.org/abs/2502.11591 https://arxiv.org/abs/2502.11591
20. Ophir, O., Shefi, O., & Lindenbaum, O. (2024). Classifying neuronal cell types based on shared electrophysiological information from humans and mice. Neuroinformatics, 22(4), 473–486. https://doi.org/10.1007/s12021-024-09675-5
21. Shadbahr, T., Roberts, M., Stanczuk, J. et al. The impact of imputation quality on machine learning classifiers for datasets with missing values. Commun Med 3, 139 (2023). https://doi.org/10.1038/s43856-023-00356-z
22. Zeng, H. (2022). What is a cell type and how to define it? Cell, 185(15), 2679–2695. https://doi.org/10.1016/j.cell.2022.06.031
23. S. Dutta, A. Arunachalam and S. Misailovic, "To Seed or Not to Seed? An Empirical Analysis of Usage of Seeds for Testing in Machine Learning Projects," 2022 IEEE Conference on Software Testing, Verification and Validation (ICST), Valencia, Spain, 2022, pp. 151-161, https://doi.org/10.1109/ICST53961.2022.00026
24. Erickson, N., Mueller, J., Shirkov, A., Zhang, H., Larroy, P., Li, M., & Smola, A. (2020). AutoGluon-Tabular: Robust and accurate AutoML for structured data. https://doi.org/10.48550/arXiv.2003.06505
25. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785
26. Y. H. Huang, Y. C. Ko, and H. C. Lu, "An optimal classification method for biological and medical data," *Mathematical Problems in Engineering*, vol. 2012, Article ID 398232, 2012. https://doi.org/10.1155/2012/398232
27. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774. https://doi.org/10.48550/arXiv.1705.07874
28. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953
29. Poole, I., Lisowska, A., Beveridge, E., & Wang, R. (2018). Classification method and apparatus (U.S. Patent No. 10,163,040 B2). U.S. Patent and Trademark Office. https://patents.google.com/patent/US10163040B2
30. Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. ACM Transactions on Knowledge Discovery from Data (TKDD), 6(4), 1–21. https://doi.org/10.1145/2382577.2382579 https://doi.org/10.1145/2382577.2382579

31. Nayak, S., & Khilar, P. M. (2024). Data imputation in healthcare applications. In AI Healthcare Applications and Security, Ethical, and Legal Considerations (pp. 49-67). IGI Global. https://doi.org/10.4018/979-8-3693-7452-8.ch004 https://doi.org/10.4018/979-8-3693-7452-8.ch004
32. Moepya, S. O., Akhoury, S. S., Nelwamondo, F. V., & Twala, B. (2015). Measuring the impact of imputation in financial fraud. In Computational Collective Intelligence: 7th International Conference, ICCCI 2015, Madrid, Spain, September 21-23, 2015, Proceedings, Part II (pp. 533-543). Springer International Publishing. https://doi.org/10.1007/978-3-319-24306-1_52
33. Brown, M.L. and Kros, J.F. (2003), "Data mining and the impact of missing data", Industrial Management & Data Systems, Vol. 103 No. 8, pp. 611-621. https://doi.org/10.1108/02635570310497657 https://doi.org/10.1108/02635570310497657
34. N. Kumar, M. A. Hoque, M. Shahjaman, S. M. S. Islam, and M. N. H. Mollah, "Metabolomic biomarker identification in presence of outliers and missing values," *BioMed Research International*, vol. 2017, Article ID 2437608, 2017. https://doi.org/10.1155/2017/2437608
35. S. Jäger, A. Allhorn, and F. Bießmann, "A benchmark for data imputation methods," *Frontiers in Big Data*, vol. 4, Article 693674, 2021. https://doi.org/10.3389/fdata.2021.693674
36. Srivastava, S., & Mishra, D. (2023). Severity of error in hierarchical datasets. Dental Science Reports, 13. https://doi.org/10.1038/s41598-023-49185-z
37. Gulati, V., & Raheja, N. (2021). Efficiency Enhancement of Machine Learning Approaches through the Impact of Preprocessing Techniques. International Conference on Signal Processing. https://doi.org/10.1109/ISPCC53510.2021.9609474

A Appendix

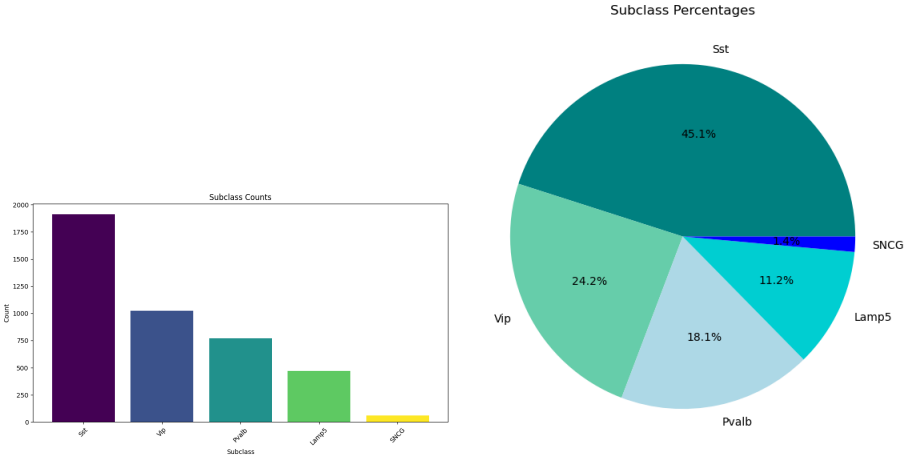


Fig. 7. Subclass Counts and Percentages

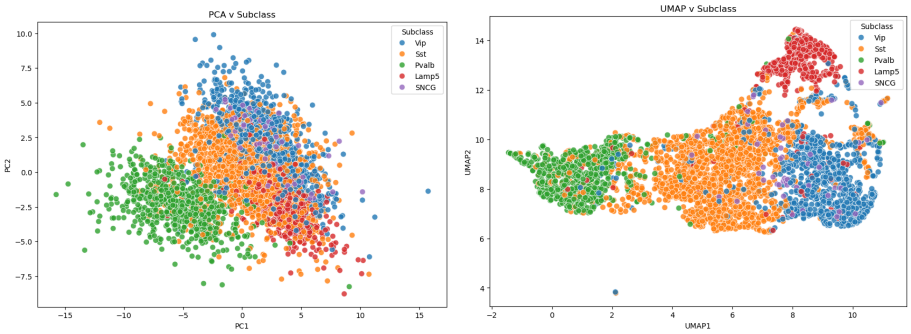


Fig. 8. PCA and UMAP

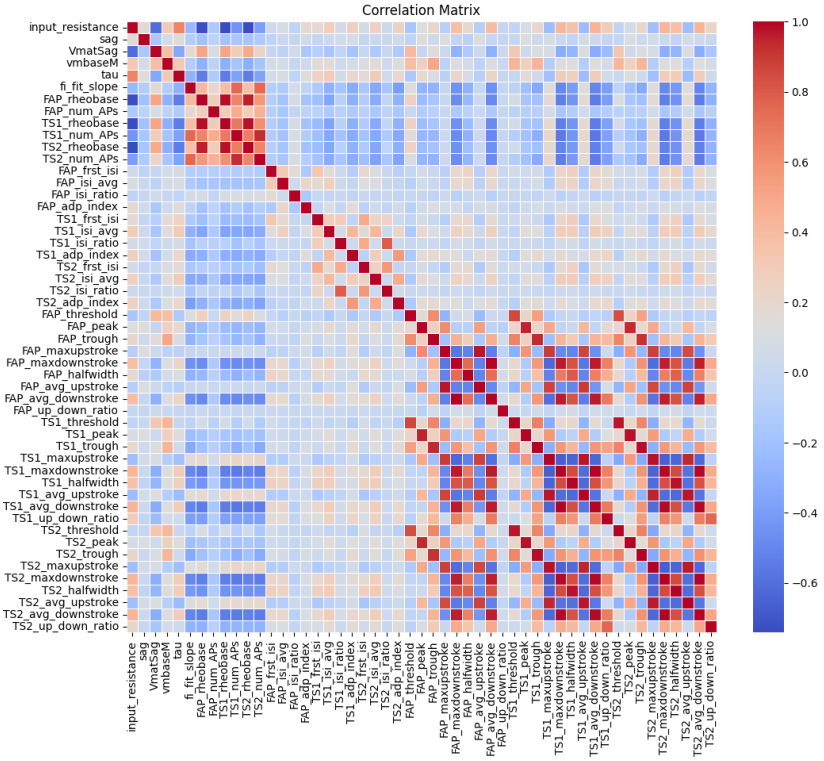


Fig. 9. Correlation Matrix

B Appendix

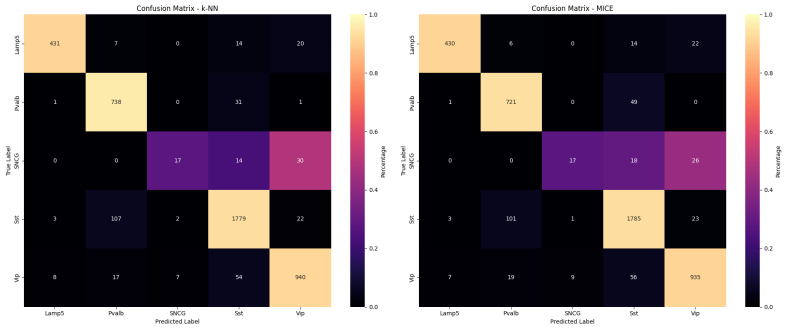


Fig. 10. Confusion Matrices - kNN (left) and MICE (right)

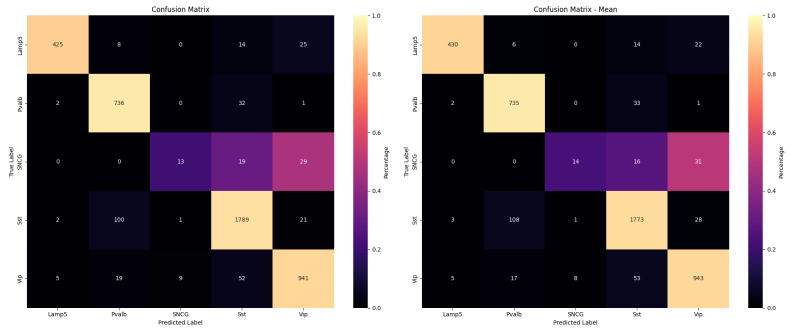


Fig. 11. Confusion Matrices - SoftImpute (left) and Mean (right)

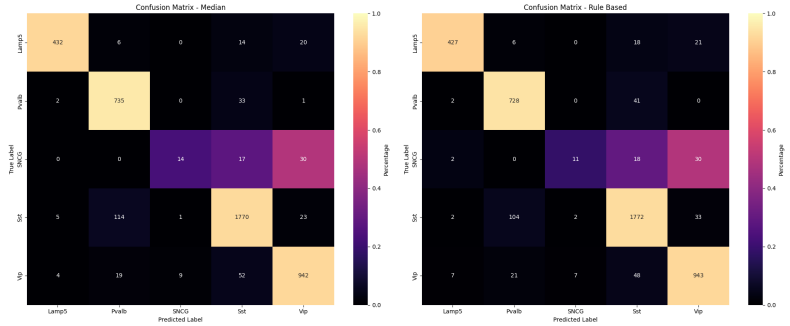


Fig. 12. Confusion Matrices - Median (left) and Rule-Based (right)

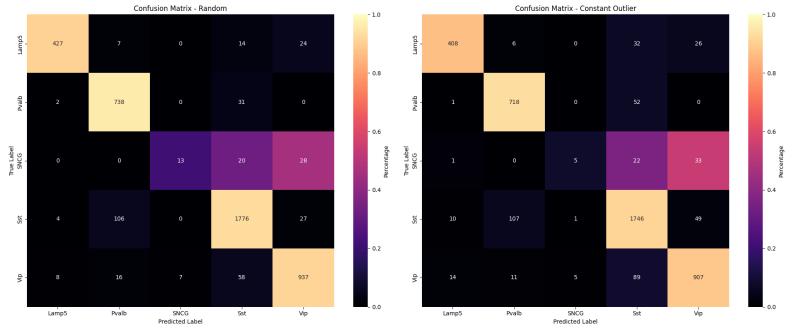


Fig. 13. Confusion Matrices - Random (left) and Constant Outlier (right)