Decoding the Divine: Religious Affiliation Prediction using Multiple Model Approaches

Caroline Delva
Courtney Green
Lizzie Healy
Sophia Rutman

DSAN 6600
Georgetown University
April 29th, 2025

ABSTRACT

This study compared three machine learning algorithms for classifying religious groups using Pew Research Center survey data. Clustering analysis revealed meaningful overlap within traditions, especially among Protestants and the Religiously Unaffiliated, highlighting the complexity of religious labels and the challenges of perfect classification. The multi-layer perceptron achieved moderate performance, with a validation accuracy of 67.7%, a precision of 65.2%, a recall of 46.1%, and an F1 score of 48.8%. The test accuracy was 68.0%, and the macro F1 score and weighted F1 score were 0.44 and 0.66, respectively. The model performed well on dominant classes but struggled with minority groups. Quadratic discriminant analysis reached 67% accuracy (weighted F1 = 0.70) for the full 13-class task and 89% (weighted F1 = 0.90) for a simplified version, though it failed to predict smaller groups (F1 = 0.00). The transformer, optimized through hyperparameter tuning, achieved the strongest results with a 94.9% test accuracy (macro F1 score = 0.56, weighted F1 score = 0.93). While it still struggled with minority classes, its self-attention mechanisms and learned embeddings enabled strong generalization across religious categories. Overall, these findings demonstrate that while machine learning models, particularly transformers, can accurately predict broad patterns of religious affiliation, capturing the full diversity and complexity of belief systems will necessitate richer data, more sophisticated modeling strategies, and enhancements in survey design.

I. Introduction

Religious belief is a deeply personal matter, often deeply rooted in one's culture and upbringing. As the world offers a diverse range of cultures and traditions, it also presents a wide range of religious beliefs. However, what constitutes a religious belief? Religious affiliations are often associated with a particular Deity, such as Jesus, Buddha, or Vishnu, and their associated customs, including pilgrimage, Eid, or Ramadan. However, the focus is often on the religion itself, rather than on the believer, leaving the profiles of each religion's followers unexplored.

The question remains: What do the followers of these religious groups actually believe? Moreover, can we distinguish between these groups based on factors such as their followers' personal demographics, religious convictions, spiritual practices, and political ideologies? Fortunately, the Pew Research Center (2024) provides large-scale survey data on religious affiliation and related social factors, offering a foundation for understanding the religious landscape in the United States.

Moreover, prior research has demonstrated the value of machine learning in uncovering patterns in religious behavior and affiliation. For example, studies using data from the World Values Survey have shown that models such as random forests can effectively identify key predictors of religiosity, including age, income, political views, and life satisfaction (Keely, 2023). These studies have demonstrated that these factors are not only crucial for understanding the religious landscape but also align with sociological theories, such as secularization and existential security, which help explain why religiosity varies across different groups (Keely, 2023).

In the United States, demographic projections suggest that Christianity is on the decline, as younger generations are moving away from traditional religious beliefs and becoming more unaffiliated. Pew Research Center data analysis indicates that if current trends continue, Christians could become a minority within the next few decades, while the unaffiliated population grows to form the largest religious category (Pew Research Center, 2022). This imbalance in affiliation across religions is highly evident in their dataset, as Christians and the unaffiliated had the most respondents, while other religions, such as Buddhism and Jehovah's Witnesses, had fewer respondents.

Building on prior studies, the current research aims to evaluate the performance of three machine learning algorithms —multi-layer perceptron (MLP), quadratic discriminant analysis (QDA), and transformers —in classifying religious affiliation using the Pew Research Center's survey dataset. It is hypothesized that the multi-layer perceptron will outperform the other models, as it is better suited to capturing the complex patterns within the data.

By doing so, the current study hopes to contribute to a deeper understanding of the followers of various religious affiliations through the application of machine learning and deep learning. This research could not only inform more effective policymaking, such as crafting policies that promote religious tolerance and social integration, but also foster a greater understanding between different religious communities, encouraging empathy and cooperation in diverse societies.

II. Data Collection

The data for this research were collected from the Pew Research Center's Religious Landscape Study. It is meant to fill in the religious data gap left by the U.S. census and to "paint a religious portrait of the United States" (Pew Research Center). The study itself is a survey

conducted utilizing Address-Based Sampling (ABS) to select participants and send invitations to the study randomly. This methodology is chosen in the hope of achieving a representative sample of the entire US population across all fifty states. Once the recipient receives the invitation, they are given the option to complete the survey online, on paper, or through a phone call. The study was conducted in three waves, in 2007, 2014, and 2023-2024. For our purposes, we elected to utilize the most recent wave of data, which was collected between July 17, 2023, and March 4, 2024. Within this wave, the Pew Center successfully surveyed 36,908 individuals. There are a total of 130 questions that span personal demographics, religious beliefs, religious practices, political views, and social values. Examples of questions include:

1. "What is the highest degree or level of school you have completed?"
2. "Do you believe in God or a universal spirit?"
3. "Outside of attending religious services, how often do you pray?"
4. "As of today, do you lean more to…The Republican Party or the Democratic Party?"
5. "More women in the workforce: Is this a change for the better, or worse, or does it not make much difference?"

Every question is answered by selecting from the specified options; there are no open-ended questions that allow respondents to express their opinions. They are, however, permitted to refuse or leave a question blank if they find the responses do not completely align with their true views. For all questions, except those about religious denomination, which is more numerous, between two and twelve options are provided for the individual.

For this research, we obtained the raw data from the Pew Research Center's website. We then removed any observations that contained missing values or instances of non-response. Additionally, a few variables were excluded from the analysis due to being restricted or unnecessary. The final dataset for this research consisted of 21,764 observations and 101 variables. Our variable of interest will be the CURREL (current religion) variable, which is a constructed variable created from the survey question that asked respondents to select their current religious denomination affiliation from over 50 options. This variable was re-grouped multiple times until the final variable here which included 13 larger religious categories:

1. Protestant
2. Catholic
3. Church of Jesus Christ of Latter-day Saints (Mormon)
4. Orthodox Christian
5. Jehovah's Witness
6. Other Christian
7. Jewish
8. Muslim
9. Buddhist
10. Hindu
11. Other World Religions
12. Other Faiths
13. Religious Unaffiliated

In all our use cases, we removed the 14th option, which was deemed refused or uninterpretable, and in some cases, we dropped the ambiguous religious categories (including 'other Christian', 'other world religions', and 'other faiths'). On one occasion, we will utilize RELTRAD as the response variable, which is derived from the same survey question and is

slightly more comprehensive, with 15 religious categories. The predictor variables comprise the remaining 100 variables, which are derived from the survey questions.

A. Class Imbalance

We note here that the CURREL variable does play host to a stark amount of class imbalance. As seen in Figure 1, the survey collects responses from many Protestants, Catholics, and religiously unaffiliated individuals. However, the remaining religious categories each have very few observations. Jehovah's Witnesses and Other World Religions are the most sparse, with only 34 and 59 instances, respectively.

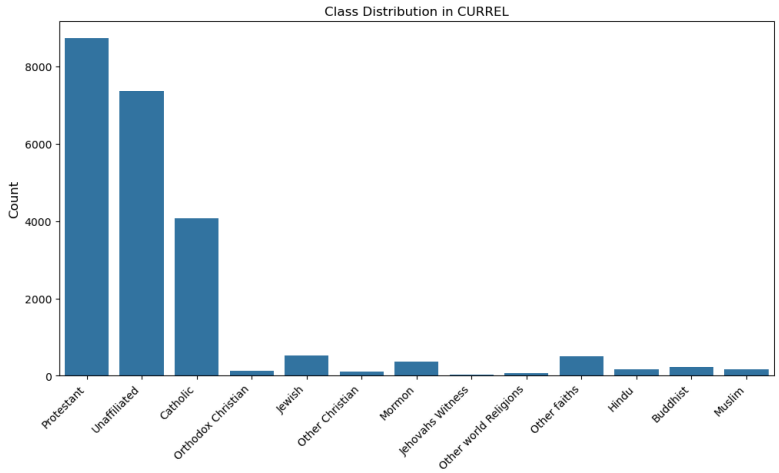


Figure 1. Class Distribution of the CURREL variable.

This is not entirely unaligned with the true distribution of religions across the United States; however, due to the inability to survey the entire population, we are left with a problematic number of observations for the minority classes and an evident class imbalance. We will, therefore, approach our analysis with consideration for the class imbalance and employ statistical methods to mitigate its impact.

III. Clustering

Before running supervised learning models, we first conducted an unsupervised clustering analysis to gain a better understanding of the underlying structure of respondents' religious, political, social, and demographic profiles. Our goal was to investigate whether individuals naturally form distinct sociocultural profiles and whether these divisions correspond to formal religious affiliations. Clustering also provided early insight into challenges for supervised classification, such as overlapping group boundaries and intra-group diversity, that might complicate predictive modeling.

A. Methods

We employed K-Means clustering, utilizing features selected through a stepwise feature selection process with a Random Forest classifier. Stepwise selection retained the 50 features that achieved the highest cross-validated accuracy on a 3-fold StratifiedKFold, balancing predictive power with dimensionality reduction. The final feature set included indicators across five primary domains: political values, religious belief and institutional participation, spiritual behavior and experiences, demographic characteristics, and worldview orientation toward

science and meaning. Political features included measures such as views on government size and support for abortion rights. Religious practices were measured by the frequency of prayer, attendance at religious services, membership in a religious congregation, participation in children's religious activities, and the importance of religion throughout different life stages. Spiritual features captured engagement in personal and communal spiritual activities (e.g., meditation, pilgrimage, visiting sacred sites), as well as beliefs related to supernatural forces, spiritual experiences, and the perceived interaction between science and spirituality. Demographic variables included race, age cohort, marital status, and migration patterns. Worldview features measured beliefs related to evolution and perceptions of science-religion conflict.

We determined the number of clusters (k) using two diagnostic techniques: the Elbow Method, which identifies diminishing returns in within-cluster sum of squares (WCSS), and Silhouette Scores, which evaluate how similar individuals are to their assigned cluster compared to other clusters. Elbow plots suggested a possible inflection around k = 5, while silhouette analysis peaked at k = 2, implying that a two-cluster solution provided the strongest average separation between groups. Based on these results, we selected two clustering solutions to explore further: k = 2 to capture broad socio-cultural divisions, and k = 5 to uncover finer-grained subgroup structures.

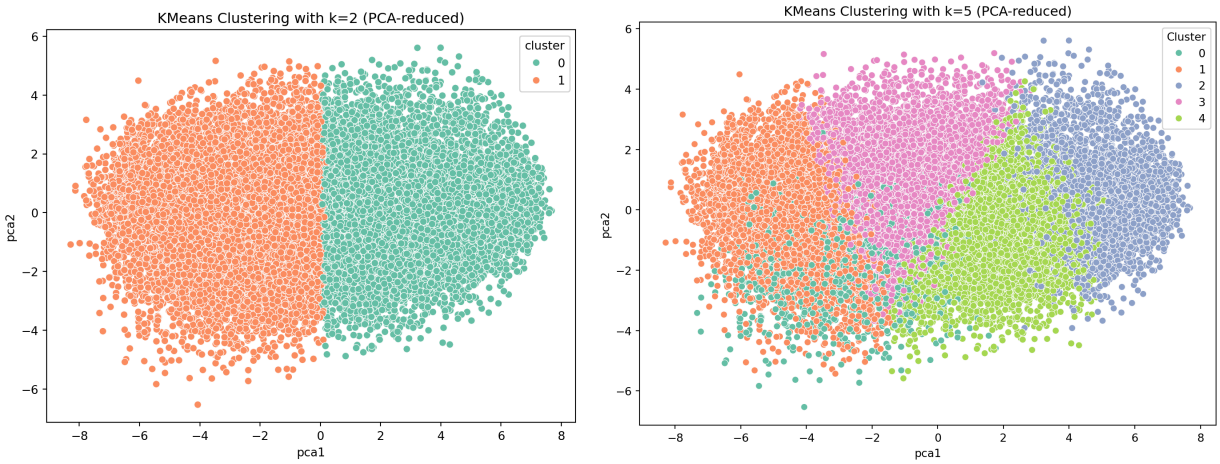


Figure 2. PCA Visualization of K-Means Clustering Solutions (k=2 and k=5) on Full Best Subset Selection Feature Set

After assigning clusters, we projected the standardized feature space into two dimensions using Principal Component Analysis (PCA). PCA reduced the high-dimensional feature space into two principal components that captured the greatest variance across respondents. We visualized the resulting component scores, coloring respondents by their cluster assignment, to assess the distinctness or overlap of the identified clusters in the reduced space.

We then cross-tabulated cluster membership against religious tradition (RELTRAD) and current religious affiliation (CURREL) to evaluate how naturally occurring socio-cultural clusters aligned, or failed to align, wth formal religious identities. We then repeated the clustering process on subsets of features within specific domains (e.g., political ideology, religious practice, spiritual activity), selecting the best number of clusters (k) separately for each subset based on silhouette scores. This allowed the grouping patterns within each domain to

emerge organically from the data, rather than pre-specifying k = 2 or k = 5, as in the full-feature clustering.

Finally, for each clustering solution, we calculated standardized feature means across clusters to profile the distinguishing characteristics of each group. We compared the distributions of religious affiliations across clusters using stacked bar plots, providing a visual sense of how religious affiliation was embedded within broader sociocultural patterns. This combined approach—feature-selected K-Means clustering, PCA visualization, cross-tabulation with religion variables, and interpretation of standardized feature profiles—offered a multi-layered understanding of the underlying structure in the data and highlighted potential challenges for religious classification using supervised learning models.

B. Results

At k = 2, clustering produced two broad socio-cultural profiles (Figure 3). The first cluster consisted of individuals who were more secular, less institutionally religious, and more exploratory in their spiritual approach. This group included a large share of the Religiously Unaffiliated, as well as Mainline Protestants and some Catholics. The second cluster reflected a more traditionally religious profile, with higher levels of formal religious practice, stronger belief in God, and more conservative social and political values. Evangelical Protestants, Historically Black Protestants, and Catholics were heavily represented. Moreover, though religious affiliation contributed to the separation between groups, it was not a perfect divider. Protestant groups, in particular, appeared across both clusters, highlighting significant ideological and behavioral diversity within formal religious traditions.

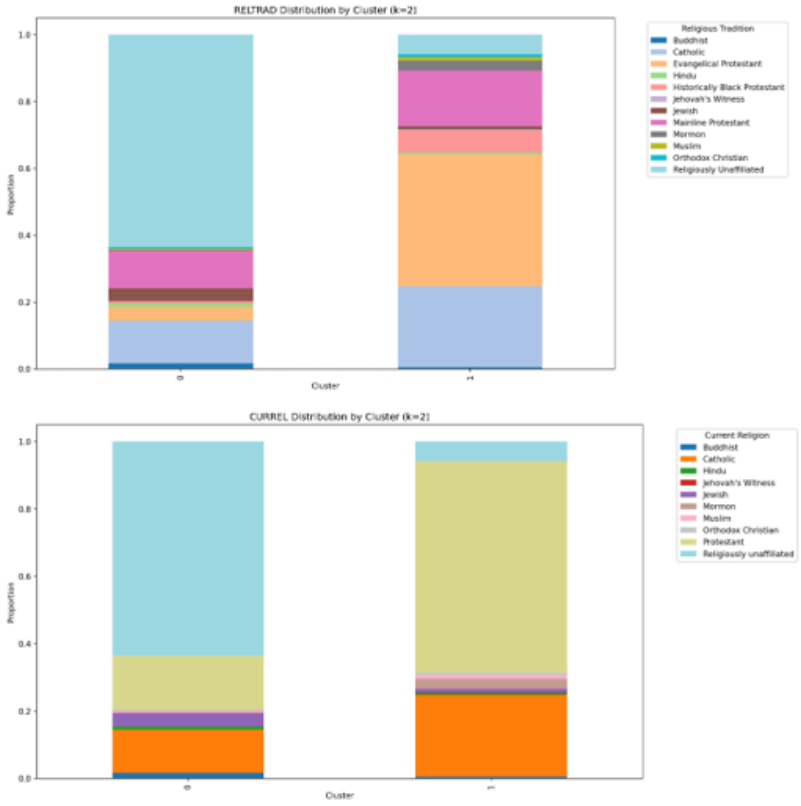


Figure 3. RELTRAD and CURREL Distributions by Cluster Assignment (k = 2)

When we expanded the analysis to k = 5, finer-grained divisions became visible (Figure 4). One cluster represented a moderate or mixed group, with centrist views on religious, political, and social issues. Another cluster was highly secular and liberal, characterized by low religiosity, strong support for diversity, and high belief in science. A third cluster comprised the most religiously devout individuals, who held strong beliefs in God and traditional family structures, and had lower endorsement of scientific reasoning. A fourth cluster reflected civic engagement and social concern, combining moderate religious practice with progressive social values. The final cluster exhibited a pattern of spiritual openness and individual exploration, characterized by moderate engagement in spiritual practices, yet skepticism toward organized religion. These clusters are reflected in the standardized feature averages shown in Figure 5.

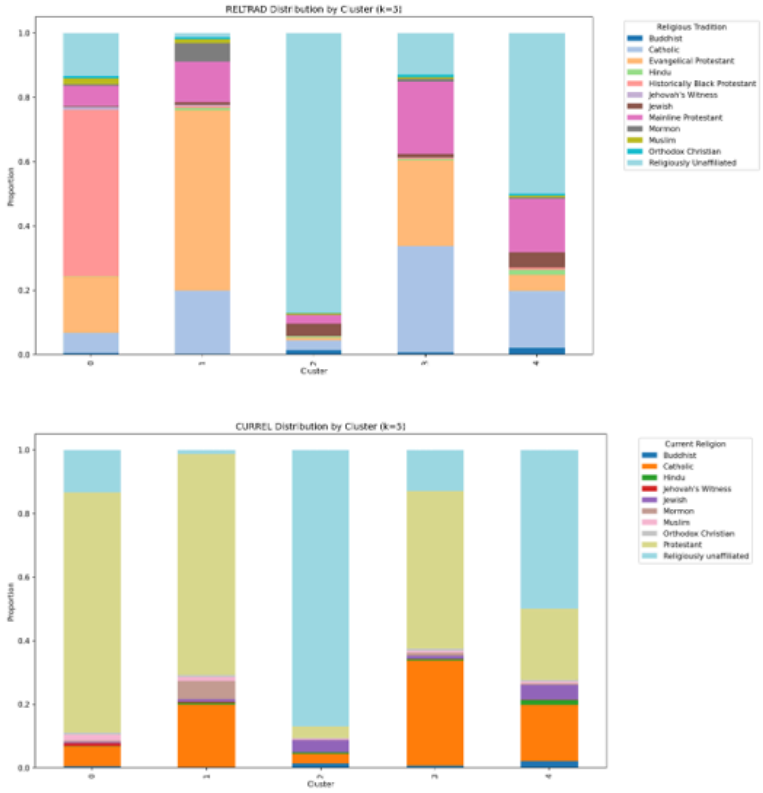


Figure 4. RELTRAD and CURREL Distributions by Cluster Assignment (k = 5)

To better understand the sources of variation, we also clustered respondents within specific feature domains (Appendix A4-A8). Clustering on political values showed that religious affiliation and political ideology were tightly linked (Appendix A4). Evangelical Protestants clustered together on the conservative side, while the Religiously Unaffiliated leaned heavily liberal. Clustering on religious practice separated highly observant individuals, often Evangelicals, from those with lower levels of formal religious participation, many of whom were unaffiliated (Appendix A5). Clustering of spiritual practices yielded a surprising result: high levels of spiritual activity were more prevalent among the Religiously Unaffiliated than among members of organized religions (Appendix A6). Demographic clustering was less distinctive (Appendix A7); younger, more racially diverse, and more mobile individuals leaned toward unaffiliated identities, while older, more stable individuals (has lived in the same place longer,

less likely to have recently moved, and more often born in the U.S.) were more often Catholic or Evangelical. Finally, clustering by science and worldview variables revealed that, although belief in evolution and scientific reasoning varied across groups, these patterns were not as sharply divided as those observed in political or religious practice domains (Appendix A8).

Overall, the clustering results indicate that while formal religious affiliation influences social and political attitudes, it does not fully account for them. Many traditions, particularly Protestantism, encompass individuals with diverse levels of religiosity, political beliefs, and social engagement. Religious identity often overlaps with political ideology, spirituality, and demographic factors, but not in a one-to-one way. This underscores a limitation in classification performance: when the same label, such as "Protestant" or "Unaffiliated," spans very different behavioral profiles, predictive accuracy may suffer. Clustering helps explain that variability — and shows why religion is a noisy target for prediction and exemplifies how unsupervised learning is a crucial complement to supervised modeling, enabling the discovery of hidden structure and the identification of label noise that can hinder classification performance.

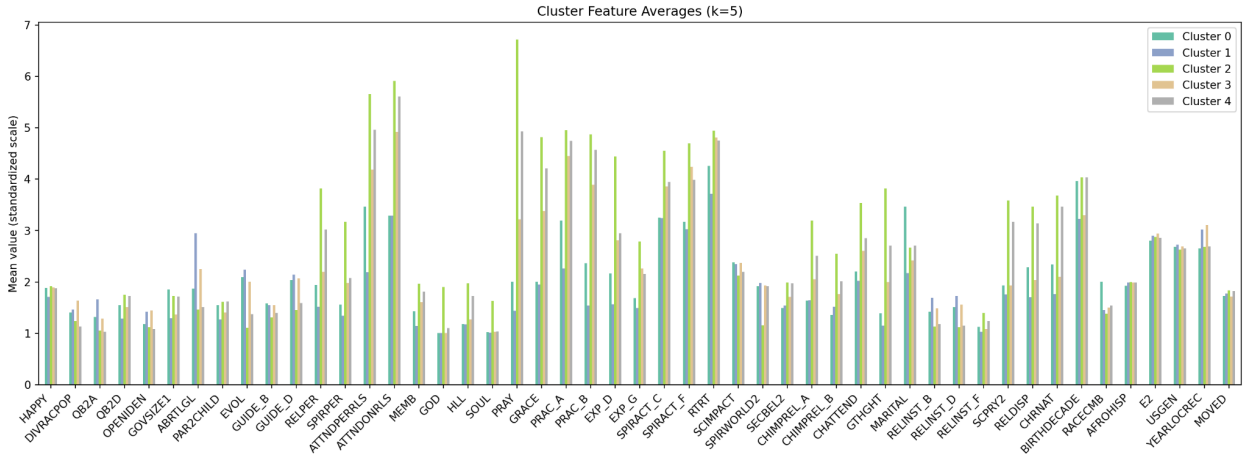


Figure 5. Standardized Feature Averages by Cluster (k = 5)

IV. Quadratic Discriminant Analysis

Quadratic Discriminant Analysis (QDA) is a statistical method used to classify data points into classes with non-linear decision boundaries. This method assumes that the data is normally distributed within each class, but allows for different covariances across classes, which enables quadratic and, therefore, increasingly complex boundaries. It utilizes a discriminant function, which is based on Bayes' Theorem and prior probabilities, to classify data points.

A. Methods

To perform QDA, we employed three different submodels with differing versions of the response variable CURREL. The first iteration included the variable in full with each of the 13 religious categories. The second iteration attempted a more straightforward approach by categorizing the religious denominations as Christian, Non-Christian, or Unaffiliated. The third and final iteration re-grouped the variable into seven categories to group the minority classes: Protestant, Unaffiliated, Catholic, Jewish, Mormon, Other, and Jehovah's Witness. Each of these iterations was meant to help understand how class imbalance may impact the results and if the outcome metrics could be improved with slightly different class groupings. Each of the

submodels initially used the full set of predictor variables and then employed only the 50 variables selected through feature selection.

For preprocessing the data for QDA, we standardized the predictor values using the built-in StandardScaler function from the scikit-learn library. This specific method performs standard z-score normalization by subtracting the mean and dividing by the standard deviation of the predictor variable across observations, resulting in a value ranging from 0 to 1. Additionally, SMOTE resampling from the Imbalanced-learn library was applied to the response variable. This is a statistical method that attempts to address class imbalance by balancing the resampling of minority classes.

The QuadraticDiscriminantAnalysis function from scikit-learn was used for the model training, and hyperparameter tuning was performed. The hyperparameter reg_param, which regularizes the per-class covariance, was tuned using an exhaustive search method. The results were inconclusive as each of the hyperparameters had minimal impact on the validation accuracy. Thus, we selected 0.2 as a reasonable default value.

B. Results

Each of the submodels was evaluated through a comprehensive classification report, a confusion matrix, and ROC curves for each class. We will primarily focus on the overall accuracy of the model and the weighted F1 score. This choice is due to accuracy being easily comparable across QDA, MLP, and transformers, and because weighted F-1 gives us a better sense of how the model is performing across all classes, not just the majority classes.

	Baseline Model		**Selected Features**	
	Accuracy	**Weighted F-1**	**Accuracy**	**Weighted F-1**
Submodel 1	0.67	0.70	0.60	0.66
Submodel 2	0.89	0.90	0.91	0.91
Submodel 3	0.68	0.71	0.65	0.69

Table 1. Accuracy and F-1 results from QDA across all submodels.

From Table 1, we observe that utilizing feature selection worsened the results for submodels 1 and 3, and made an extremely minimal difference for submodel 2. Therefore, we focus on the results from the baseline model with the complete set of predictor variables.

For submodel 1, which included all 13 religious categories, we observed an accuracy of 0.67 and a weighted F-1 score of 0.70, suggesting that of all the classification predictions the submodel made across the classes, 67% correctly matched the true religious label for the particular data point. The F-1 score gives an average of the precision (true positive predictions over positive predictions) and recall (true positives over true positive and false negative), and weights the values based on the sample size of each of the classes. Taken together, these metrics made the submodel look as though it was performing decently well; however, when we investigated the classification report, we saw that even the weighted F-1 is not able to account for the extreme class imbalance. In this case, the Protestant class had an F-1 score of 0.73, and the Unaffiliated class had 0.86, however, this is opposed to Other Christian, Orthodox Christian,

and Other World Religions, which had a precision, recall, and F-1 score of 0.00 (Appendix A1). Figure 6 depicts the ROC curves for the two most prominent classes and the single least prominent. This begins to highlight how the QDA model can correctly classify the majority classes and is able to perform exceptionally well (0.97 AUC score for Unaffiliated); however, it struggles with the minority classes (0.89 AUC for Other World Religions).

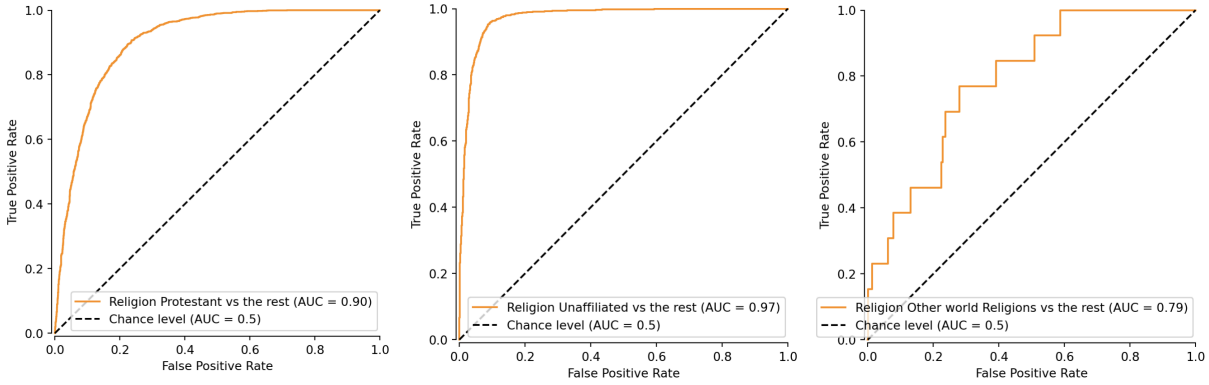


Figure 6. ROC Curves for the Protestant, Unaffiliated, and Other World Religions. These are the two most represented religions and the least represented religion, respectively.

Submodel 2, a simplified version that only requires classification into Christian, Non-Christian, and Unaffiliated categories, showed improved performance. The accuracy score of 0.89 means that, among all predictions, the submodel was able to classify between the three religions 89% of the time. The weighted F-1 score of 0.90 also indicates a strong classifier submodel even across the classes. However, this again is hiding some of the story. In this submodel, there are 13,442 Christians, 7,355 Religiously Unaffiliated, and 1,651 Non-Christians; thus, while the class imbalance is improved, it is not altogether solved. The full classification report showed an F-1 score of 0.95, 0.90, and 0.46, respectively. The Non-Christian class has a precision score of 0.38 and a recall of 0.57, meaning it is struggling to identify Non-Christians correctly and to avoid misclassifying other observations as Non-Christian. Again, the results are not consistent across the classes, as Christian has a precision of 0.97 and a recall of 0.92 (Appendix A2). While this submodel is clearly an improved classifier as opposed to submodel 1, it is still being hampered by the issue of class imbalance.

Submodel 3, while an attempt to improve upon submodel 1, performs only marginally better. It had an accuracy of 0.68 and a weighted F-1 of 0.71. Following the theme of class imbalance, unaffiliated gave an F-1 score of 0.86 and Protestant 0.73, which Mormon contrasted with 0.29 and Christian with 0.15 (Appendix A3). The sole improvement between the two submodels was that this was able to avoid any precision and recall scores of 0.00.

Overall, the only submodel of QDA that showed signs of classification ability was submodel 2, which only produced simplified results between Christian, Non-Christian, and Unaffiliated. However, this submodel, along with the other two, struggled heavily with the issue of class imbalance, making QDA inapplicable for this research scenario.

V. Multi-Layer Perceptron

A Multi-Layer Perceptron (MLP) is a type of feedforward artificial neural network that processes weighted inputs with biases through one or more hidden layers, applying activation functions at each layer to generate a predicted output. (Hickman, 2025). The current study

utilizes an MLP model to classify individuals into 12 religious groups, such as Protestant, Catholic, Muslim, Buddhist, and others, based on their responses to the Pew Research survey.

Its architecture expects an input layer that accepts 50 features selected through the top 50 stepwise feature selection. The input is followed by two fully connected hidden layers, each containing 64 neurons. The neurons of the hidden layers are passed through nonlinear activation functions to either Tanh or Leaky ReLU, which are tunable hyperparameters. The model optionally regularizes or drops its large weights through either L1 or L2 regularization or dropout.

The final layer is a Dense output layer with 12 neurons, which are the number of classes predicted during training and are passed through a Softmax activation to produce a probability distribution over the predicted religious classes. The model uses sparse categorical cross-entropy as its loss function because the targets were label encoded.

A. Methodology

The data was normalized using StandardScaler normalization so that features with larger scales do not dominate the learning process. Next, the target labels, which were the religious affiliation categories, were mapped to integers using a LabelEncoder, resulting in 12 distinct classes.

The data was first separated into three sets: training, validation, and test. Stepwise (forward-floating) feature selection using an off-the-shelf Scikit-Learn Random Forest classifier as the base estimator was conducted to determine the best features for the model to learn from. Features were added and removed based on their contribution to model accuracy during 3-fold stratified cross-validation. Ultimately, the top 50 features were selected to assist the model in classifying religious affiliation. As can be seen in Figure 7, the top contributing features were the importance respondents attributed to religion, the extent to which they considered themselves religious, the significance of the Bible to them, the frequency of their prayer, and the importance they attached to attending services in person.

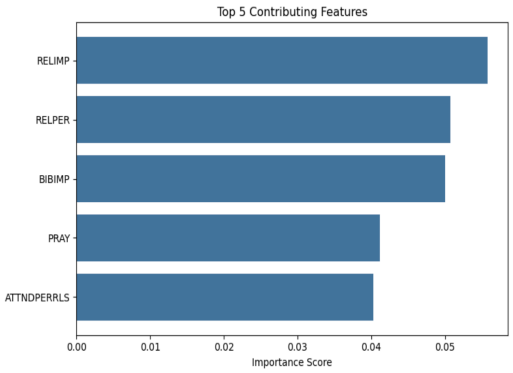


Figure 7. Top 5 Contributing Features of Feature Selection

The MLP model was subsequently trained using early stopping with a patience of 10 epochs, in order to stop training if the validation loss failed to improve over 10 consecutive epochs. The model was trained using a batch size of 16, allowing it to update its weights after every 16 training samples. The maximum number of epochs was set at 80, which was rarely reached as early stopping prevented the model from overfitting.

To train the best-performing model, a grid search over the following hyperparameters was conducted on the validation set:

- Optimizers: rmsprop, sgd, adam
- Activation functions: ReLU, Tanh, Leaky_ReLU
- Regularization types: None, L1, L2
- Regularization strengths: 0.001, 0.01, 0.1
- Dropout rates: 0, 0.2, 0.5

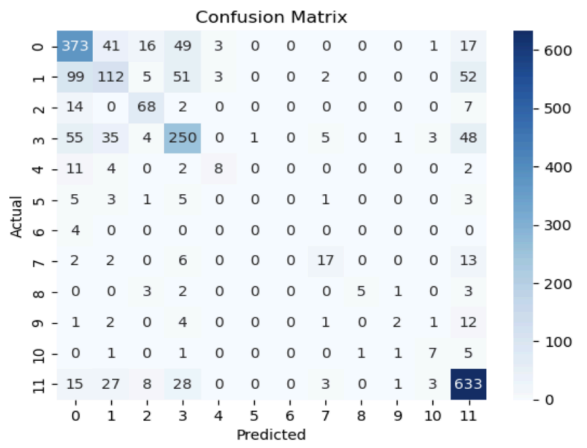
The performance of each set was evaluated, and the accuracy, precision, recall, and F1 scores were reported.

B. Results

The best configuration was the SGD optimizer, Tanh activation function, L1 regularization with a strength of 0.001, and no dropout, based on the hyperparameter tuning step. This configuration achieved a validation accuracy of 67.7%, meaning that 67.7% of the validation set was correctly predicted in the multiclass classification task. Additional performance metrics included a precision of 65.2%, a recall of 46.1%, and an F1 score of 48.8%, indicating a moderate balance between false positives and false negatives.

When evaluated on the unseen test set, the model achieved an accuracy of 68%, a macro F1 score of 0.44, and a weighted F1 score of 0.66. As the macro F1 score is the unweighted average across all classes, it demonstrates that the model was able to predict the rare classes 44% of the time correctly. However, the higher weighted F1 score reflects stronger performance on the majority classes, as it accounts for the class imbalance.

As seen in Figure 8., when taking a closer look into the model's performance, the classification report and confusion matrix MLP demonstrated that the model performed the best on dominant religious groups such as Protestant and Religiously Unaffiliated but produced low recall for minority classes which included Orthodox Christian, Jehovah's Witness, or Buddhist.



Classification Report (Test Set):

	precision	recall	f1-score	support
1100	0.64	0.75	0.69	500
1200	0.49	0.35	0.41	324
1300	0.65	0.75	0.69	91
10000	0.62	0.62	0.62	402
20000	0.57	0.30	0.39	27
30000	0.00	0.00	0.00	18
40001	0.00	0.00	0.00	4
50000	0.59	0.42	0.49	40
60000	0.83	0.36	0.50	14
70000	0.33	0.09	0.14	23
80000	0.47	0.44	0.45	16
100000	0.80	0.88	0.84	718
accuracy			0.68	2177
macro avg	0.50	0.41	0.44	2177
weighted avg	0.66	0.68	0.66	2177

Figure 8. Multi-layer Perceptron Confusion Matrix and Classification Report for Test Set Performance

The model's performance disparity can be attributed to the class imbalance noted within the data, as some classes had greater representations across all sets as compared to others. While regularization and dropout might have helped prevent the model from overfitting, they could not

fully compensate for the lack of representation in the minority classes. Future work could explore techniques such as class weighting, data augmentation, or engineering to address this class limitation and improve minority class performance.

VI. Transformer

A. Methods

We created two transformer models, each with the same architecture. Their only difference came with the number of classes they attempted to classify. The first model examined only eleven different religions, while the second model examined fourteen. In the first classification model, we eliminated the more ambiguous categories, hoping to boost accuracy rates. The classes eliminated included other world religions, other Christianity, and other faiths.

The architecture is similar to that of a traditional transformer built for NLP tasks. However, we did not include a decoder in our architecture. Decoders are used to generate content, which is not necessary for a classification task. However, we did include other traditional elements.

Each algorithm begins by projecting numerical features into the embedding space, ensuring that the encoder can complete the self-attention and embedding algorithms to pass the information forward. The number of encore blocks depends on the optimal number determined by hyperparameter tuning, but each block follows a consistent structure. They use multi-head self-attention, where the number of heads also depends on hyperparameter tuning. The data then passes through two layers of normalization, a feedforward network, and a dropout layer to prevent overfitting. The dimensions of the feedforward network, as well as the dropout rates, depend on hyperparameter tuning once again. After the data passes through the encoder, it is processed by a final fully connected layer, which outputs a vector of size either ten or thirteen, depending on the model, with logits associated with each element.

To train the model, we used categorical cross-entropy as the loss function and a warm-up scheduler to gradually adjust the loss with each epoch.

We completed hyperparameter tuning to optimize each model. We specifically optimized the embedding dimension, feedforward dimension, number of attention heads, number of encoder blocks, dropout rates, and the learning rate. A specific combination of these features, different for each of the two models, did prove successful in increasing the accuracy rating.

B. Results

Through hyperparameter tuning, we achieved impressive accuracy results. In the model identifying eleven religions, the optimal hyperparameters are as follows:

Embedding Dimension	128
Feedforward Dimension	256
Attention Heads	4
Encoder Blocks	2
Dropout	0.3
Learning Rate	0.0001

Table 2: Ten-class Model Optimal Hyperparameter Selection

These hyperparameters led to a test accuracy rate of 76.07%. The training of this optimal model is visualized below.

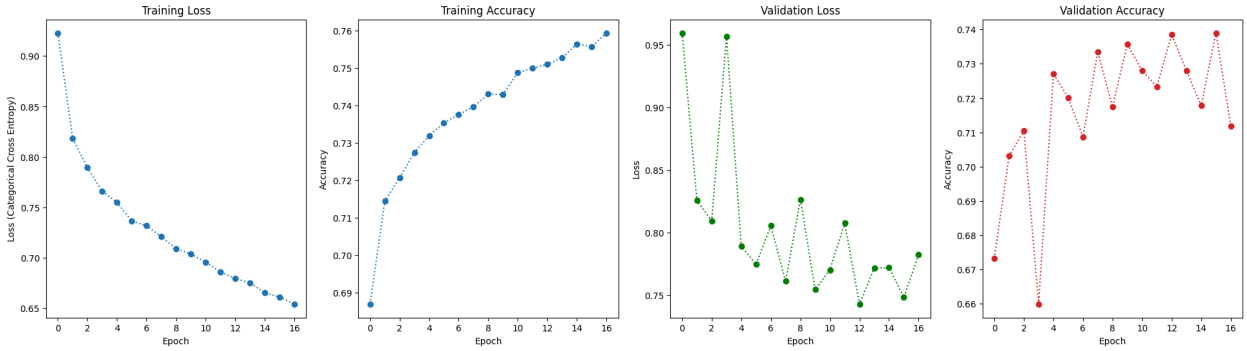


Figure 9. Ten-class Transformer Training Epochs

The confusion matrix is presented in Figure 10.

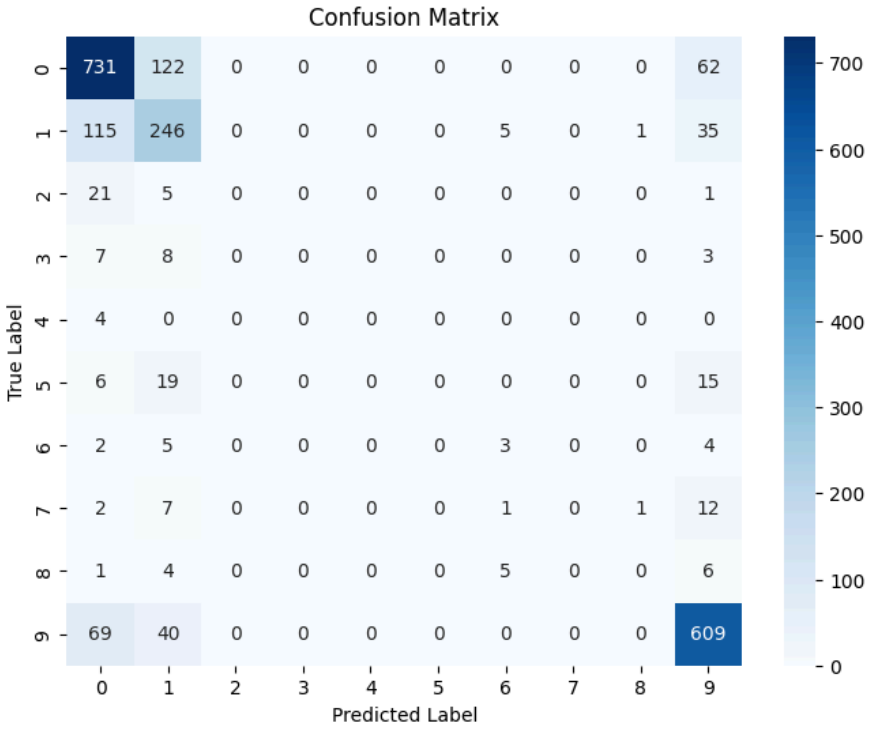


Figure 10. Ten-class Transformer Confusion Matrix

The confusion matrix clearly demonstrates the differing proportion of true positives. For example, while class 0 (Protestant) had precision, recall, and F1 scores of 0.7780, 0.8273, and 0.8019, respectively, these numbers dipped as low as 0.5714, 0.1739, and 0.2667 for other, less numerous classes. This indicates that while the model was fairly confident and accurate when predicting the largest class, its ability to both correctly identify and correctly predict smaller classes was much weaker. The low recall for these minority classes suggests that the model

frequently failed to detect them, and the low precision means that when it did predict these classes, it was often wrong.

The macro F1 score for this test set was 0.5148, and the weighted was 0.7472, which highlights the imbalance across classes. A macro F1 score averages performance across all classes equally, and its relatively low value reflects poor model performance on rare classes. In contrast, the higher-weighted F1 score, which accounts for class frequency, indicates that the model performed better on the more populated classes, and that these larger groups disproportionately influenced the overall performance metrics.

Class three was an outlier in that the precision, recall, and F1 were all zero. Although this did not significantly impact the weighted F1 score due to its low sample size, it does indicate that the transformer is still struggling with classes of low size. In fact, multiple classes did not present a single true positive, highlighting the strong effect of the dominant classes on the final accuracy rating. It was able to predict these numerous classes well, but essentially failed to recognize the less-represented groups.

However, this is still extremely promising, as the overall accuracy here is still higher than that achieved by other models, likely due to the self-attention mechanisms it incorporates, which are absent in the other models. The embeddings ensure that the final feedforward dimension can properly assign logits to all classes, no matter the number of observations in that class.

We achieved a significantly higher accuracy rating using fourteen classes, including all religions represented in the survey. The optimal hyperparameters are presented below:

Embedding Dimension	128
Feedforward Dimension	128
Attention Heads	2
Encoder Blocks	2
Dropout	0.1
Learning Rate	0.00005

Table 3: Thirteen-class Model Optimal Hyperparameter Selection

The training plots are presented in Figure 11.

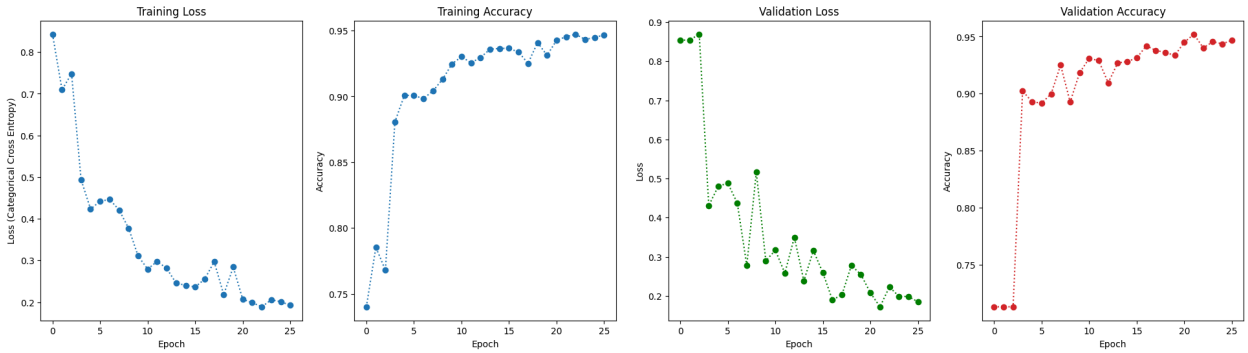


Figure 11. Thirteen-class Transformer Training Epochs

The final validation accuracy of the training was much higher than that of the eleven classes, which was extremely promising. The final test accuracy was 94.86%, which was staggering compared to the accuracy of the previous transformer model, as well as clustering, QDA, and MLP. This highlights the transformer's strong generalization ability, even when faced with a more complex, multi-class prediction task.

The F1 score was higher across samples than in the previous classification model, with a macro-averaged F1 score of 0.5637 and a weighted score of 0.9282. The increase in weighted F1 suggests that the model maintained very strong performance on the larger classes, reflecting its ability to capitalize on abundant data where available. However, the gap between macro and weighted F1 scores indicates that performance was still highly uneven across classes, with weaker performance on a few smaller groups. Although the transformer significantly improved overall predictive power, it continued to struggle with some rare classes, a common challenge in imbalanced datasets.

It was able to perfectly predict multiple classes, including classes 0, 1, 2, and 3. This is impressive, especially for classes 2 and 3, as they contain very few data samples. The model did not incorrectly predict any of these classes as another class, indicating that the embeddings effectively captured the characteristics of these survey responses in a meaningful way. This model was able to handle the class imbalance in these classes much more gracefully than any model presented above.

The high accuracy level was primarily driven by the model's success in predicting the more numerous classes. The transformer was able to correctly predict an extremely high number of Protestant, Unaffiliated, and Catholic samples, while, similar to the model above, there were still classes that had precision, F1, and recall scores of zero due to a total absence of correct predictions. These classes are clearly visible in Figure 12 below, highlighting that some minority classes remain challenging to classify correctly under the current data distribution.

Unfortunately, no samples from class 13 were included in the test set due to the nature of the random split, which introduces further difficulty in evaluating the model's ability to generalize to that class. This issue highlights an important avenue for future studies: ensuring equivalent splits across datasets will provide a more accurate and fair evaluation of model performance, particularly for rare categories.

Compared to traditional models, the transformer not only delivered higher overall accuracy but also demonstrated robustness to class imbalance and better scalability to a larger number of classes. This suggests that the self-attention mechanism and learned embeddings together provide a significant advantage in capturing nuanced patterns within the survey

responses. Figure 12 further demonstrates the success of the transformer, particularly in its ability to model dominant classes with high reliability while maintaining overall competitive performance.

An avenue for future work is exploring why the fourteen-class model performed better than the eleven-class model. It is possible that having more classes helped the model learn more distinct patterns and relationships in the data. The larger number of classes may have allowed the model to separate the categories better, while the eleven-class setup could have created more confusion between certain groups. Further analysis of the model's attention and embeddings could help explain this difference.

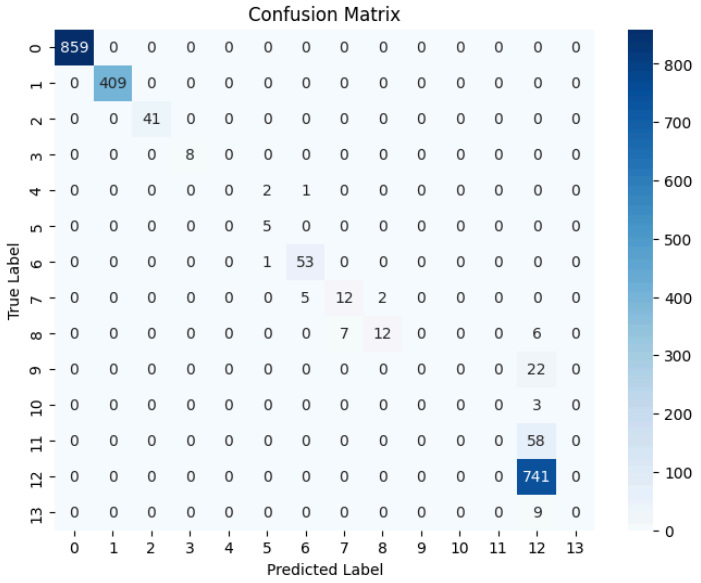


Figure 12. Thirteen-class Transformer Confusion Matrix

VII. Discussion

Our study had a few limitations. First, as we discussed previously, the class imbalance is presented in Figure 1. The transformer was able to look past this imbalance, but QDA and our MLP were unable to do so. The models simply did not have enough data to train on when looking at classes other than Protestant, Religiously Unaffiliated, and Catholic. Even with the use of SMOTE resampling, the models successfully identified other classes from the test set. The transformer, despite achieving a high overall accuracy rate, was unable to identify a significant portion of the data upon testing properly. This points to a need for a change in survey distribution.

In addition, survey data is notorious for bias. Users were filling out this data based on self-reporting, which can introduce inaccuracies due to social desirability bias, misunderstanding of questions, or intentional misreporting. This potential for response bias may have impacted the quality and consistency of the training data, and thus the models' ability to generalize. Moreover, certain demographic groups may have been underrepresented in the survey responses, limiting the model's exposure to the full diversity of the population.

VIII. Conclusion

Our study examined the challenges of and opportunities in predicting religious affiliation using both supervised and unsupervised learning methods. While clustering revealed meaningful socio-cultural divisions, it also showed that religious identity is not a simple, one-dimensional target. Significant ideological, behavioral, and demographic diversity exists within traditions like Protestantism and the Religiously Unaffiliated, making religion a noisy label for supervised models. Among classifiers, the transformer model achieved the highest predictive performance, especially when classes were simplified. However, all models struggled to fully capture the complexity of minority religious groups, emphasizing that class imbalance and internal diversity remain serious obstacles.

Several directions could strengthen future work and address these challenges more effectively. While we used basic SMOTE resampling to oversample minority classes, more advanced techniques, such as SMOTE-ENN, could be more effective. SMOTE-ENN first creates synthetic examples to balance the class distribution and then removes data points, both real and synthetic, that are noisy or likely to be misclassified (Batista et al., 2004). This two-step process balances the dataset while cleaning up ambiguous cases that could confuse the model, which is particularly important in our study, where categories such as Hindus, Buddhists, and Jehovah's Witnesses have very few observations and considerable overlap with broader social patterns. Using SMOTE-ENN could help the model better distinguish between these smaller religious groups without overfitting to noisy or marginal cases, leading to improved recall and F1 scores for underrepresented traditions (Batista et al., 2004).

Moreover, a key future direction involves enhancing the survey by incorporating open-ended questions. Survey methodology researchers Schonlau and Couper argue that, while easier to process, fixed-choice questions often fail to capture the complexity of respondents' true beliefs and experiences (Schonlau & Couper, 2016). In our context of religious affiliation, this limitation likely contributed to the label noise and ambiguity we observed and allowing respondents to describe their religious identity in their own words could capture nuances such as mixed traditions, non-institutional spirituality, or culturally religious identities that structured categories like "Protestant" or "Unaffiliated" cannot fully reflect. Although open-text data would require additional processing effort, semi-automated text-mining methods exist to manage this at scale (Schonlau & Couper, 2016). However, more significantly, these types of improvements to the survey would produce a richer, more authentic dataset, ultimately supporting better clustering and classification by reducing the noise inherent in rigid, predefined labels.

Another improvement is to shift from single-label to multi-label classification, as studies have shown that multi-label modeling, particularly using transformer-based architectures like BERT, can successfully capture the complexity of survey responses where individuals align with multiple categories simultaneously (Schonlau et al., 2023). In our context, allowing respondents to hold multiple affiliations or predicting probabilities across traditions better reflects the reality of belief systems where cultural, spiritual, and formal identities overlap. This could help avoid forcing individuals into rigid single categories like "Protestant" or "Unaffiliated" when their behaviors and beliefs span multiple traditions, improving both predictive accuracy and conceptual validity. Without adjustments to both data collection and model design, religious affiliation is likely to remain a noisy and challenging target for supervised learning.

Works Cited

1. Batista, Gustavo & Prati, Ronaldo & Monard, Maria-Carolina. (2004). A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. SIGKDD Explorations. 6. 20-29. 10.1145/1007730.1007735.
2. Hickman, James. *DSAN 6600 Class Lecture, Topic: "Dense Feed-Forward Fully Connected Neural Network."* Georgetown University, DSAN 6600, April 28, 2025.
3. Hickman, James. *Course Notes on Clustering Overview.* Georgetown University, DSAN 5000, accessed 2025.
4. Jafarigol, E., Keely, W., Hortag, T., Welborn, T., Hekmatpour, P., & Trafalis, T. B. (2023). Religious Affiliation in the Twenty-First Century: A Machine Learning Perspective on the World Value Survey [Preprint]. arXiv. 2310.1087
5. Pew Research Center. (2022). Modeling the future of religion in America. Pew Research Center.
6. Pew Research Center. (2025). *2023–24 U.S. Religious Landscape Study Interactive Database*. https://www.pewresearch.org/religious-landscape-study/
7. Schonlau, M., & Couper, M. P. (2016). Semi-automated categorization of open-ended questions. *Survey Research Methods*, *10*(2), 143–152. https://doi.org/10.18148/srm/2016.v10i2.6213
8. Schonlau, Matthias & Weiß, Julia & Marquardt, Jan. (2023). Multi-label classification of open-ended questions with BERT. 10.48550/arXiv.2304.02945.

Appendix

	precision	recall	f1-score	support
Buddhist	0.22	0.18	0.20	39
Catholic	0.56	0.65	0.60	815
Hindu	0.48	0.41	0.44	34
Jehovahs Witness	1.00	0.10	0.18	10
Jewish	0.23	0.49	0.31	107
Mormon	0.18	0.65	0.28	69
Muslim	0.53	0.27	0.36	37
Orthodox Christian	0.00	0.00	0.00	27
Other Christian	0.00	0.00	0.00	26
Other faiths	0.14	0.38	0.20	87
Other world Religions	0.00	0.00	0.00	13
Protestant	0.82	0.66	0.73	1777
Unaffiliated	0.91	0.81	0.86	1449
accuracy			0.67	4490
macro avg	0.39	0.35	0.32	4490
weighted avg	0.74	0.67	0.70	4490

0.6746102449888641

CONFUSION MATRIX:

```
[[   7    1    2    0    3    0    2    0    0    6    0    2   16]
 [   4  530    2    0   42   39    0    3    4    7    1  171   12]
 [   5    2   14    0    1    0    3    0    0    0    0    1    8]
 [   0    1    0    1    0    0    0    0    0    0    0    8    0]
 [   0   17    0    0   52    1    0    1    0   10    0    7   19]
 [   0   12    0    0    0   45    0    0    0    0    0   11    1]
 [   3    8    4    0    2    0   10    1    0    0    0    6    3]
 [   1   10    1    0    1    0    1    0    0    0    0   12    1]
 [   0    6    0    0    1    1    0    0    0    8    0    7    3]
 [   2    3    0    0    2    0    0    0    0   33    0    5   42]
 [   1    0    1    0    0    0    1    0    0    4    0    2    4]
 [   2  342    0    0   74  161    1    6   11    7    0 1165    8]
 [   7   21    5    0   46    0    1    1    2  168    1   25 1172]]
```

A1. Full Classification Report and Confusion Matrix for QDA submodel 1.

CLASSIFICATION REPORT:

	precision	recall	f1-score	support
Christian	0.97	0.92	0.95	2724
Non-Christian	0.38	0.57	0.46	317
Unaffiliated	0.91	0.89	0.90	1449
accuracy			0.89	4490
macro avg	0.75	0.80	0.77	4490
weighted avg	0.91	0.89	0.90	4490

0.889532293986637

CONFUSION MATRIX:

```
[[2519  175   30]
 [  37  181   99]
 [  36  119 1294]]
```

A2. Full Classification Report and Confusion Matrix for QDA submodel 2.

CLASSIFICATION REPORT:

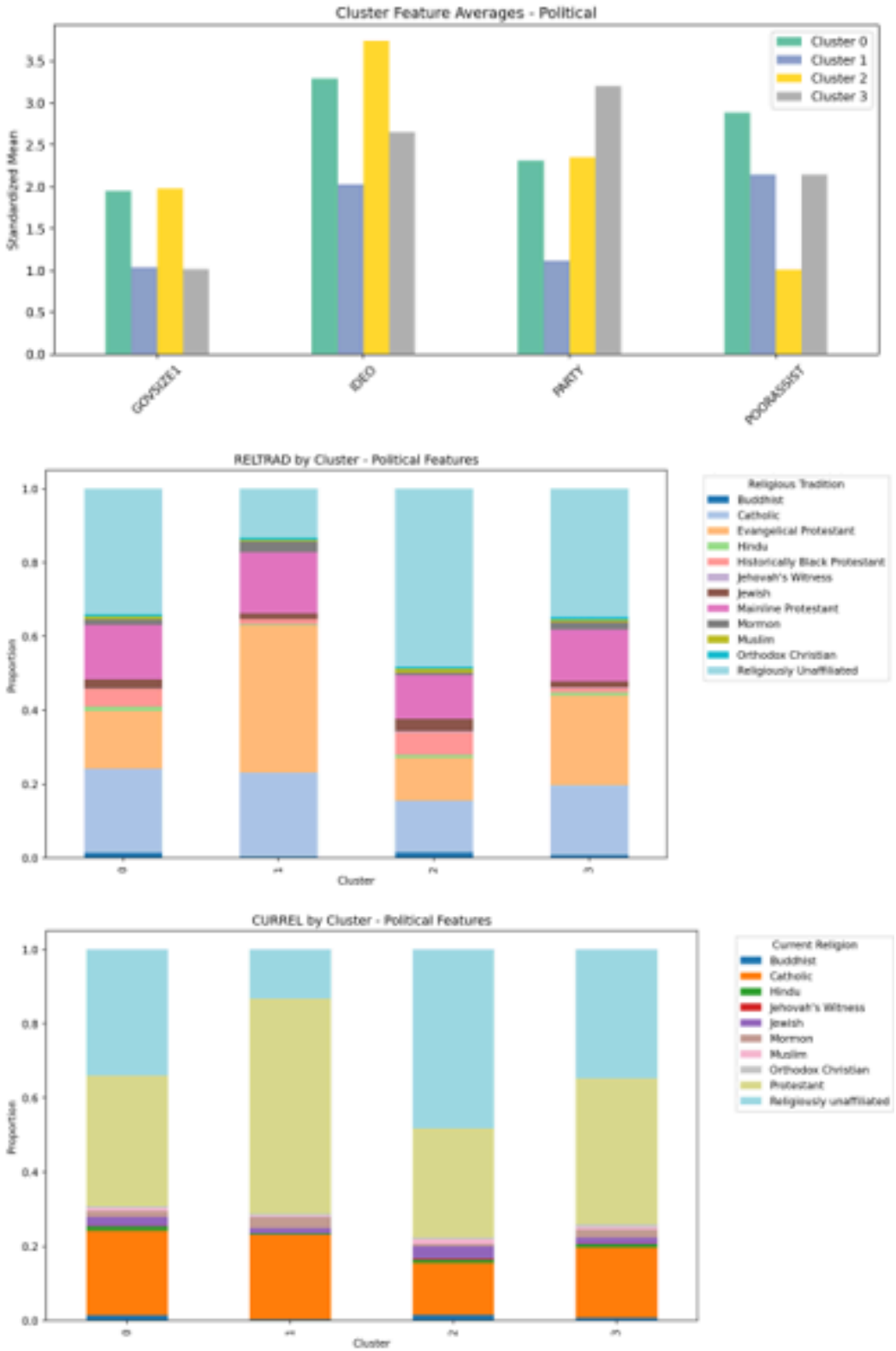
	precision	recall	f1-score	support
Catholic	0.56	0.64	0.60	815
Christian	0.14	0.16	0.15	63
Jewish	0.23	0.50	0.32	107
Mormon	0.18	0.65	0.29	69
Muslim	0.53	0.22	0.31	37
Other Religion	0.29	0.48	0.36	173
Protestant	0.83	0.65	0.73	1777
Unaffiliated	0.91	0.82	0.86	1449
accuracy			0.68	4490
macro avg	0.46	0.52	0.45	4490
weighted avg	0.75	0.68	0.71	4490

0.6841870824053452

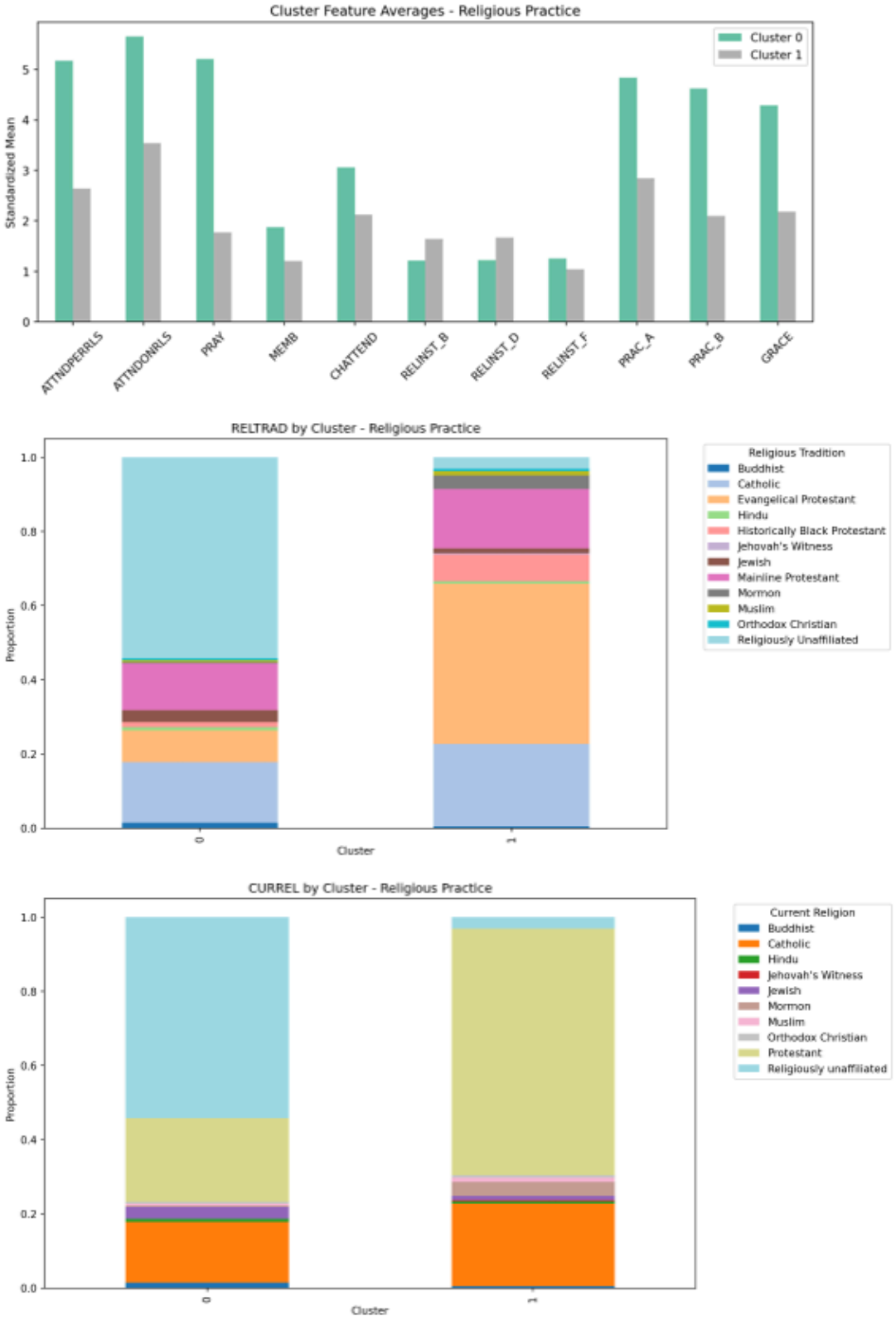
CONFUSION MATRIX:

```
[[ 525   18   43   39    0   13  166   11]
 [  14   10    4    1    1    8   22    3]
 [  18    1   54    1    0    5    7   21]
 [  12    0    0   45    0    0   11    1]
 [   6    2    1    0    8   12    6    2]
 [   5    1    8    0    4   83    7   65]
 [ 341   31   71  159    1    6 1160    8]
 [  19    8   51    0    1  159   24 1187]]
```

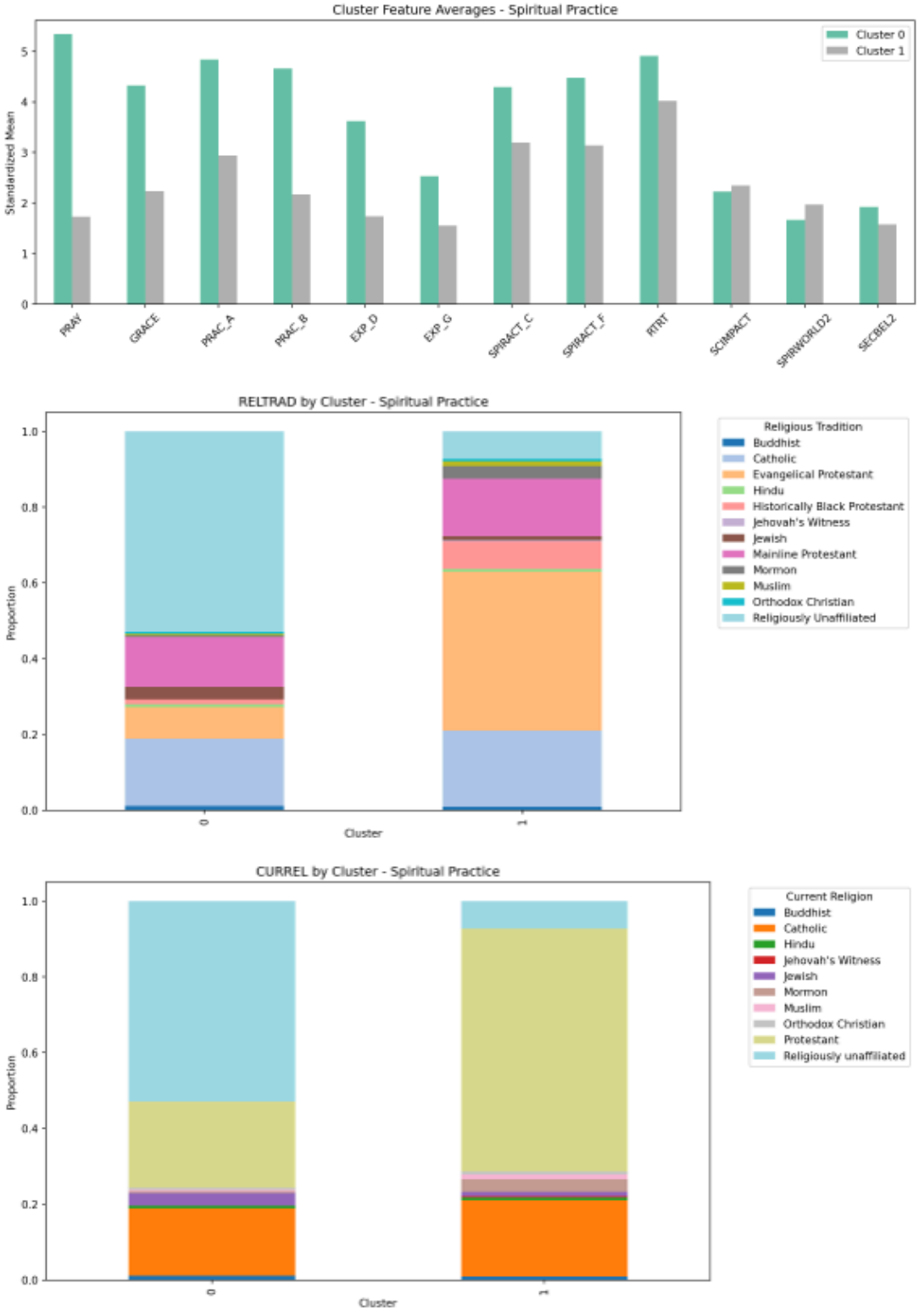
A3. Full Classification Report and Confusion Matrix for QDA submodel 3.



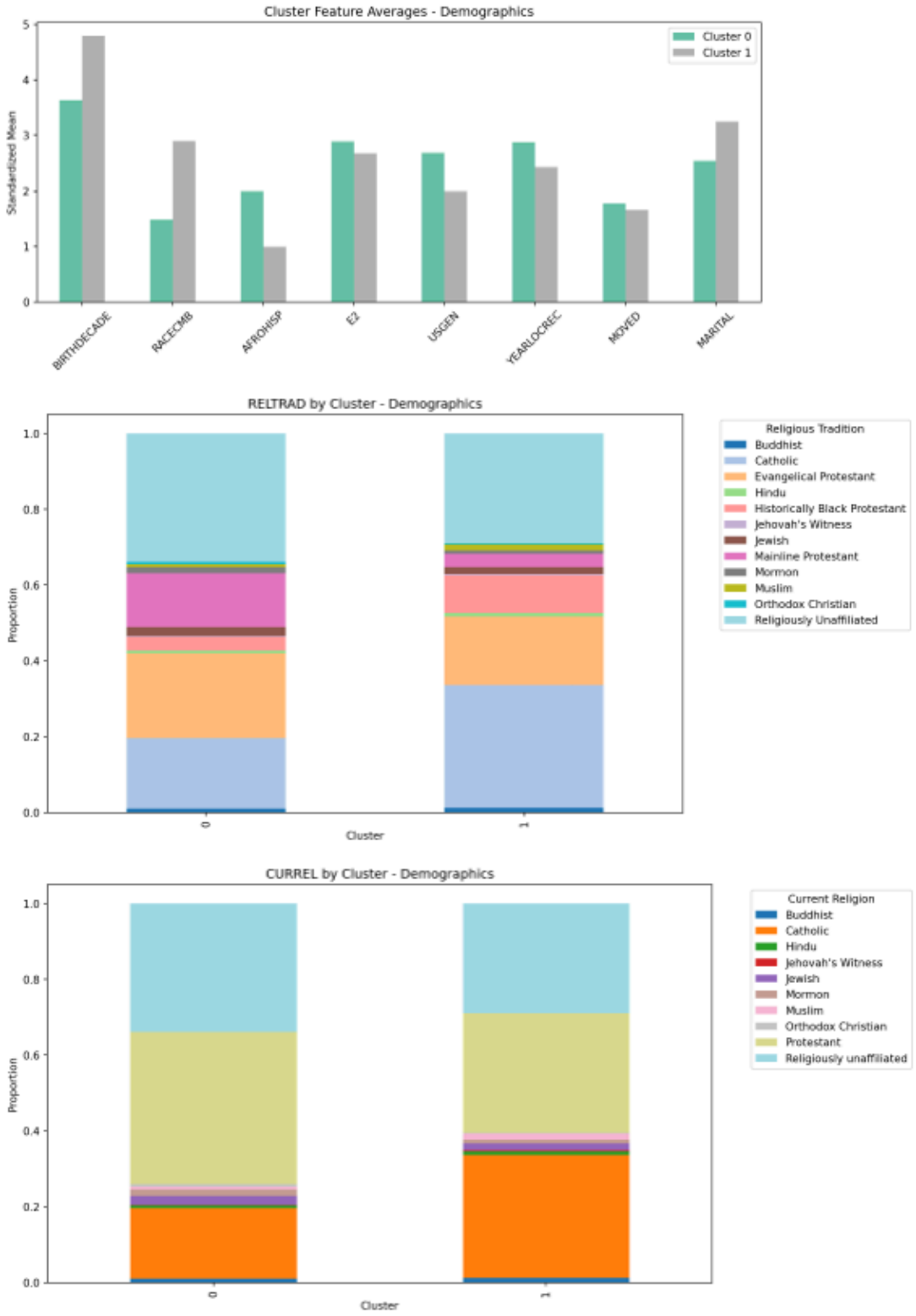
A4. Cluster Profiles and Religious Distributions Based on Political Features



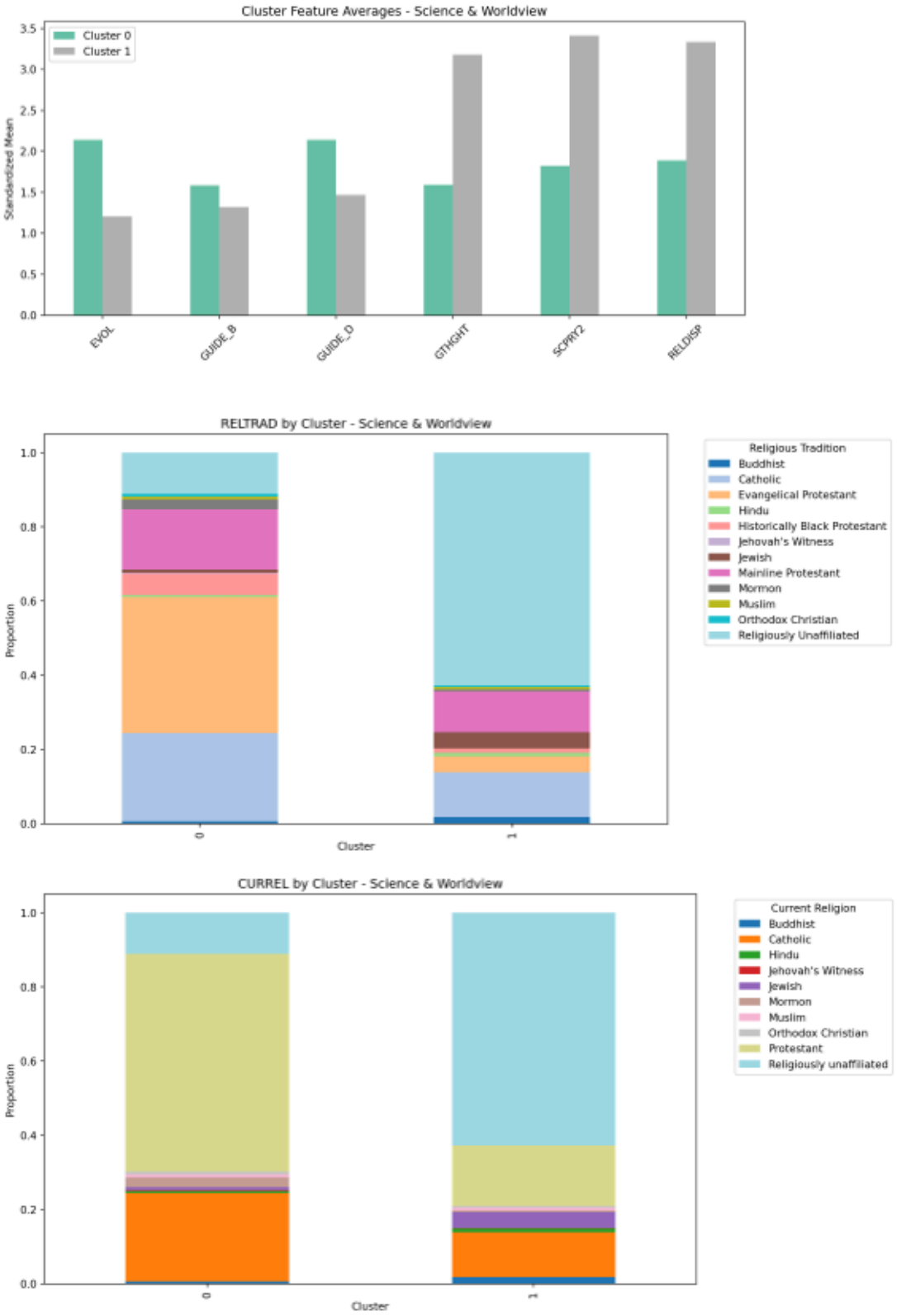
A5. Cluster Profiles and Religious Distributions Based on Religious Practice



A6. Cluster Profiles and Religious Distributions Based on Spiritual Practice



A7. Cluster Profiles and Religious Distributions Based on Demographic Features



A8. Cluster Profiles and Religious Distributions Based on Science & Worldview

Evaluation Metrics:

	precision	recall	f1-score	support
0	0.7780	0.8273	0.8019	915
1	0.7270	0.5299	0.6129	402
2	0.7143	0.3704	0.4878	27
3	0.0000	0.0000	0.0000	18
4	1.0000	0.5000	0.6667	4
5	0.5862	0.4250	0.4928	40
6	0.7143	0.3571	0.4762	14
7	0.5714	0.1739	0.2667	23
8	0.8571	0.3750	0.5217	16
10	0.7598	0.8942	0.8215	718
accuracy			0.7607	2177
macro avg	0.6708	0.4453	0.5148	2177
weighted avg	0.7502	0.7607	0.7472	2177

A9: Summary of the Ten-class transformer

Evaluation Metrics:

	precision	recall	f1-score	support
0	1.0000	1.0000	1.0000	859
1	1.0000	1.0000	1.0000	409
2	1.0000	1.0000	1.0000	41
3	1.0000	1.0000	1.0000	8
4	0.0000	0.0000	0.0000	3
5	0.6250	1.0000	0.7692	5
6	0.8983	0.9815	0.9381	54
7	0.6316	0.6316	0.6316	19
8	0.8571	0.4800	0.6154	25
9	0.0000	0.0000	0.0000	22
10	0.0000	0.0000	0.0000	3
11	0.0000	0.0000	0.0000	58
12	0.8832	1.0000	0.9380	741
13	0.0000	0.0000	0.0000	9
accuracy			0.9486	2256
macro avg	0.5639	0.5781	0.5637	2256
weighted avg	0.9116	0.9486	0.9282	2256

A10: Summary of the Thirteen-class transformer