Spring 5-14-2023

Feature Selection from Clinical Surveys Using Semantic Textual Similarity

Benjamin Warner

Recommended Citation

Warner, Benjamin, "Feature Selection from Clinical Surveys Using Semantic Textual Similarity" (2023). *McKelvey School of Engineering Theses & Dissertations*. 823.
https://openscholarship.wustl.edu/eng_etds/823

WASHINGTON UNIVERSITY IN ST. LOUIS

McKelvey School of Engineering
Department of Computer Science & Engineering

Thesis Examination Committee:
Chenyang Lu, Chair
Simon Haroutounian
Thomas Kannampallil
Cynthia Ma

Feature Selection from Clinical Surveys Using Semantic Textual Similarity

by
Benjamin C. Warner

A thesis presented to
the McKelvey School of Engineering
of Washington University in
partial fulfillment of the
requirements for the degree
of Master of Science

May 2023
St. Louis, Missouri

Table of Contents

List of Figures . . . . **iii**

List of Tables . . . . **iv**

Acknowledgments . . . . **v**

Abstract . . . . **viii**

Chapter 1: Introduction **1**

1.1 Persistent Post-Surgical Pain and Survey Data 1

1.2 Feature Selection 2

1.3 Large Language Models 3

1.4 Proposed Algorithm 4

Chapter 2: Methods **6**

2.1 Data Characteristics 6

2.2 Data & Model Preparation 7

2.3 mRMR with mutual information (mRMR-i) 9

2.4 mRMR-s and mRMR-h 11

Chapter 3: Results **13**

3.1 Model Performance 13

3.1.1 Selected Features 13

3.2 Hyperparameter Performance 15

3.2.1 Performance Over N For mRMR-i and mRMR-s 15

3.2.2 Performance Over N, α For mRMR-h 16

Chapter 4: Discussion **22**

4.1 Model Performance 22

4.2 Future Work 23

References **25**

Appendix A Feature Importance **28**

List of Figures

Figure 3.1: Heatmap of feature-feature and feature-target mutual information (MI) scores. 15

Figure 3.2: Heatmap of feature-feature and feature-target semantic textual similarity (STS) scores. 16

Figure 3.3: Performance of a Gaussian Naïve Bayes (Gaussian NB) classifier over N using mRMR-i 17

Figure 3.4: Performance of a Gaussian NB classifier over N using mRMR-i 18

Figure 3.5: Performance of an XGBoost classifier over N using mRMR-s 19

Figure 3.6: Performance of an XGBoost classifier over N using mRMR-s 20

Figure 3.7: Test and train AUROC performance over the hyperparameter space for mRMR with hybrid mutual information and semantic textual similarity (mRMR-h) using Gaussian NB, as well as the difference between them. . 21

List of Tables

Table 2.1: Demographics of the partial P5 dataset. 7

Table 2.2: Questions used to determine the label of each survey data example. . . . 8

Table 3.1: Results from using selected feature selection methods, all feature selection methods set to select up to 40 features. Best metrics for each model type and among minimal-redundancy-maximal-relevance (mRMR) are bolded. 14

Table A.1: Feature importance using Gaussian NB for mRMR with mutual information (mRMR-i) with 40 features 32

Table A.2: Feature importance using Gaussian NB using mRMR with semantic textual similarity (mRMR-s) with 40 features. 36

Table A.3: Feature importance using Gaussian NB using mRMR-h with 40 features. Features are selected with $\alpha = 0.0923671$, which was found through 5-fold flat cross-validation. 41

Acknowledgments

This thesis would not have been possible without the support of Professors Chenyang Lu and Thomas Kannampallil, who have been instrumental in supporting me and giving me opportunities to engage in impactful research the past few years. In addition, I would like to thank Professor Simon Haroutounian for making the overarching P5 research project that this thesis is a part of possible. Finally, I would like to thank Professor Cynthia Ma for contributing her time to be on my committee and providing valuable feedback.

Thanks is also in order for the two other P5 project members I worked with during this time, namely Ziqi Xu and Zihan Cao, who were helpful in providing the initial work on the dataset used in this thesis. I would also like to thank some of the other people I have worked with over the past few years in the CPSL/I^2, including Professor Joanna Abraham, Sunny Lou, Hanyang Liu, Derek Harford, Seunghwan Kim, Laura Baratta and Chioma Odozor. I have really enjoyed getting to apply new ideas at the intersection of computation and medicine with them.

I've also met a lot of great people throughout my time here at Washington University in St. Louis, thanks to communities like Bear Nation Varsity Band, WURocketry, the Dual Degree Engineering community, and so on. These past three years would not have been as fun without them. I'd also like to thank some of the faculty here, including Professors James Orr, Ed Richter, Steve Cole, Bill Siever, Sankalp Bhan, Vladimir Kurenok, Marion Neumann, and Jinsong Zhang.

I'd also like to thank some of the people in the Pacific Lutheran University for the experience I had during my time there, including Professors Sean O'Neill, Renzhi Cao, David Wolff, Ken Blaha and Laurie Murphy as well as University Congregation, Regents' Investment Fund, and the PLU Pep Band.

I also would like to acknowledge my parents and my grandparents for their immense support in enabling me to attend college, which I would have been unable to do without them. Finally, I'd like to thank my sister Rachel and my rabbits, S'mores and Puff, for being there for me.

This work was made possible through the support of the Department of Defense Congressionally Directed Medical Research Programs award W81XWH-21-1-0736.

Benjamin C. Warner

Washington University in St. Louis
May 2023

To S'mores and Puff

ABSTRACT OF THE THESIS

Feature Selection from Clinical Surveys Using Semantic Textual Similarity

by

Benjamin C. Warner

Master of Science in Computer Engineering

Washington University in St. Louis, 2023

Professor Chenyang Lu, Chair

Survey data collected from human subjects can contain a high number of features while having a comparatively low quantity of examples. Machine learning models that attempt to predict outcomes from survey data under these conditions can overfit and result in poor generalizability. One remedy to this issue is feature selection, which attempts to select an optimal subset of features to learn upon. A relatively unexplored source of information in the feature selection process is the usage of textual names of features, which may be semantically indicative of which features are relevant to a target outcome. The relationships between feature names and target names can be evaluated using large language models (LLMs) such as ClinicalBERT to produce STS scores, which can then be used to select features. This thesis introduces two new variations upon the minimal-redundancy-maximal-relevance (mRMR) algorithm that integrate semantic textual similarity (STS) into selection. The performance of STS as a feature selection metric is evaluated against preliminary survey data collected as a part of a clinical study on persistent post-surgical pain (PPSP). The results suggest that features selected with STS can result in higher performance models compared to those with the baseline mRMR algorithm.

Chapter 1

Introduction

This chapter begins with a discussion of the clinical phenomenon of persistent post-surgical pain and the collection of survey data. A review of feature selection and large language models follows, and is then finished with a brief description of the proposed solution and contributions of this thesis.

1.1 Persistent Post-Surgical Pain and Survey Data

Persistent post-surgical pain (PPSP) is the phenomenon of a patient experiencing surgically-related pain for a longer duration of time than expected [1]. Because the causes are presently unclear [2], a machine learning (ML) approach may lead to further insights into not only understanding the cause of PPSP, but also being able to predict PPSP.

One useful source of data available for predicting PPSP is survey data collected from participants, and different surveys have been designed for the purpose of capturing different characteristics of PPSP. One popular tool for obtaining data from multiple surveys is the Research Electronic Data Capture (REDCap) system [3], which has been popular in biomedical research. The results from these surveys are conglomerated together and can easily contain hundreds of features.

Since PPSP the causes of PPSP are currently unclear, a series of standard questionnaires are collected using REDCap to assess possible causes. Among several different technical issues that need to be addressed for building a ML model with this dataset is the high dimensionality of the data, a problem which is exacerbated by the relatively small number of examples upon which to train.

1.2 Feature Selection

Fitting high-dimensional data is particularly difficult when the number of examples is low—as is with clinical data collected from human subject—since a model can easily overfit on the training data. To counter this, we can employ the strategy of *feature selection*, where a subset of the overall features in a dataset are selected for learning.

Feature selection methods can be divided into three categories: *embedded*, *wrapper*, and *filter* methods. Embedded methods incorporate feature selection as a part of training, while wrapper methods interact in a feedback loop with the learning model. Filter methods select a subset of features based on properties of the dataset before the model is able to learn on the dataset, which differs from embedded and wrapper methods in that they do not form a feedback loop with the model [4]. Because of their independence, they tend to have good generalization abilities [5].

A literature review suggests that feature selection methods for survey data shows a diverse array of feature selection methods. A study examining autism spectrum disorder (ASD) survey data, examined feature selection using principal component analysis, t-distributed stochastic neighbor embedding, and denoising autoencoders; and also found that survey features targeting ASD tend to have high levels of redundancy [6]. Some of the other feature selection methods found for models involve questionnaires include wrapper model based on random forests [7], bootstrapped feature selection [8], principal component analysis, multi-cluster feature selection [9], permutation importance [10], and ReliefF [11].

One particularly important feature selection is minimal-redundancy-maximal-relevance (mRMR), which aims to maximize the *relevance* of features to the target, while minimizing the *redundancy* between selected features. This is particularly useful when we have a small number of features that are correlated and want to ensure a model incorporates as broad as a set of information as possible. The objective function of mRMR, seen in equation 1.1, is simply the the difference of relevance and redundancy, as seen in 1.2 and 1.3, respectively [12].

$$\max \Phi(D, R) = D - R \tag{1.1}$$

$$D = \frac{1}{|S|} \sum_{x_i \in S} I(x_i; c) \tag{1.2}$$

$$R = \frac{1}{|S|^2} \sum_{x_i, x_j \in S} I(x_i, x_j) \tag{1.3}$$

Because the solution space for selected features is a powerset of the possible features, we cannot directly test all solutions, and use an incremental solution to find the set of selected features. Equation 1.4 gives the solution for each candidate feature $\{X - S_{n-1}\}$ from the set of features X and the set of previously selected features at step S_{n-1} [12], [13]. This results in an algorithm with one hyperparameter, which is the number of features N to be selected.

$$\max_{x_j \in X - S_{n-1}} \left[I(x_j; c) - \frac{1}{m-1} \sum_{x_i \in S_{m-1}} I(x_j; x_i) \right] \tag{1.4}$$

Underpinning the mRMR objective function is MI between classes and features, which is defined in equation 1.5 using densities f and marginal densities f_x, f_y.

$$I(X, Y) = \iint dxdy f(x, y) \log \frac{f(x, y)}{f_x(x) f_y(y)} \tag{1.5}$$

Calculating true MI between two features is computationally costly, but can be approximated using one of several methods. The MI approximation methods used here are from scikit-learn [14], and are a synthesis of the k-nearest neighbors approaches described in [15], [16].

1.3 Large Language Models

LLMs are a class of language models that is loosely defined as having on the order of high millions or more of parameters, and have been typically built with the transformer architecture [17]. LLMs have demonstrated capabilities at many reasoning tasks involving semantic meaning [18], [19], and are highly applicable to survey data due to their text-based nature.

The first stage in training BERT is *pre-training*, where the model is trained on two unsupervised learning tasks, and then *fine-tuning*, where the model is subsequently applied to a supervised learning task [20]. Pre-training is particularly useful since it results in better generalization [21], and because it means that computationally expensive pre-trained models can be reused for different tasks [22].

Among the many fine-tuning tasks is that of STS. In this task, a model attempts to evaluate the semantic similarity of two sentences using some metric. One variation of model suited to this is the siamese neural network (SNN), where two weight-sharing neural networks generate embeddings for two input sentences, and then have their similarity computed with a function. *Cosine similarity*, as seen in equation 1.6, is one typical function used to compute the distance between embeddings [23], [24].

$$\cos(E_1, E_2) = \frac{E_1 \cdot E_2}{||E_1||||E_2||} \tag{1.6}$$

Clinical language involves vocabulary and semantic meaning that is often not present in non-clinical texts, and various pre-trained architectures exist to fill this gap. Among them is ClinicalBERT, which is an extension of the BERT model trained upon clinical notes from the MIMIC-III dataset [25], [26]. ClinicalBERT can out-perform other BERT models on tasks specific to the clinical domain [25], and can do so more efficiently than a general-purpose BERT model [27].

1.4 Proposed Algorithm

As a form of tabular data, survey answers are the result of questions that have text that may be semantically related to a target outcome, as well as semantically similar or dissimilar to one another. Intuition suggests that MI and MI are useful analogues to one another, as they both capture relationships between data statistically and semantically, respectively. STS scores derived from LLMs may then be useful for determining which questions are *relevant* to predicting a target question based, and moreover, may be useful in determining which questions are *redundant* to each other. This may be particularly true for smaller datasets where there is a limited amount of information immediately available to learn from. If we treat STS as a stand-in or compliment to MI in mRMR, then there are two potentially useful new algorithms for selecting features.

There appears to be nearly no literature examining the usage of embeddings of feature names to select features, and none examining that between given feature and target questions. The closest match examined the usage of `word2vec` continuous bag-of-words embeddings [28] trained upon Twitter data to select Google search query trends using the embeddings of a target concept [29]. The algorithm proposed in this paper differs in several different ways, with the principal difference being the proposed algorithm utilizes STS selects a combination of features maximizing equation 1.1, whereas [29] apply a one standard deviation threshold of scores for feature selection. Another major difference is the usage of ClinicalBERT to calculate scores, which is a more recent model than `word2vec`, and experimentally performs better than `word2vec` on clinical natural language processing (NLP) tasks [30]. Finally, the selection of target embeddings is differrent. In [29], they are selected purely as an adjustable hyperparameter, and for this algorithm, the embeddings are of four target questions that are defined in the survey, as well as defined label name.

The contributions of this paper are as follows:

- An examination of the role of the efficacy of utilizing STS scores generated by LLMs between feature and target questions, specifically pre-trained for a clinical context, in feature selection.
- Evaluating the performance of two novel variations of the mRMR feature selection model: mRMR-s and mRMR-h, which utilize STS as as a direct replacement and compliment for MI respectively.
- Evaluating how mRMR-s and mRMR-h can help to prevent overfitting on small survey datasets.

Chapter 2

Methods

This chapter begins with an examination of the characteristics of the dataset used, followed by the techniques used to prepare the dataset for learning. Then the baseline mRMR with mutual information (mRMR-i) algorithm is discussed, followed by a discussion of the design and implementation of the mRMR with semantic textual similarity (mRMR-s) and mRMR with hybrid mutual information and semantic textual similarity (mRMR-h) algorithms.

2.1 Data Characteristics

The data is collected from participants from the *P5 - Personalized Prediction of Postsurgical Pain* study (IRB #202101123). The participants in this study are drawn from a partially complete set of patients in the Washington University/BJC HealthCare system.

A total of 12 surveys were assigned to individual users through the REDCap system. The principal survey is the Washington University PPSP Questionnaire [1], which contains the four target outcome questions, which are described in more detail in Table 2.2. The other surveys include measures of psychological and physical pain and correlated measures.

The dataset was assembled from REDCap on February 6th, 2023. A total of 617 participants have been collected from a final goal of 2,000 participants from the WU/BJC system. Table 2.1 outlines the key characteristics of the dataset, including number of examples and general demographics.

Name	Value
PPSP Characteristics	
Individuals with Complete Mark	617
PPSP (+)	97
PPSP (-)	592
Race	
Caucasian	497
American Indian / Alaskan Native	7
Asian	4
Black / African Heritage	98
Hawaiian Native / Other Pacific Islander	1
Other	9
Prefer not to answer	7
Sex assigned at birth	
Female	425
Male	185
Age	
Age (min)	19
Age (mean)	52.4686
Age (std. dev.)	13.5219
Age (max)	75

Table 2.1: Demographics of the partial P5 dataset.

2.2 Data & Model Preparation

Several steps are taken to prepare the survey data for fitting upon the candidate ML models.

The first step taken is to prepare the label from this particular survey dataset. The label is derived from the four questions shown in Table 2.2 is determined using the binary formula $y_1 = (Q_1 \wedge Q_2) \wedge (Q_3 \geq 3 \vee Q_4 \geq 3)$. Once the labels are computed, these questions are dropped from the dataset. Since we are attempting to predict PPSP, we filter out examples that where a column indicating six-month completion has a null value.

#	Question Text	Type
y_1	`persistent_pain`	N/A
Q_1	In the past week, did you have any pain in your surgical incision or in the area related to your surgery?	Yes/No
Q_2	For pain in the area related to your surgery, did the pain start or worsen after the surgery?	Yes/No
Q_3	On a scale of zero to ten, with zero being no pain and ten being the worst pain, please fill in your average pain level during the past week, while you were at rest.	0-10
Q_4	On a scale of zero to ten, with zero being no pain and ten being the worst pain, please fill in your average pain level during the past week, when you were active or moving.	0-10

Table 2.2: Questions used to determine the label of each survey data example.

We then filter out features from a list of features pre-determined not to be relevant to the prediction of PPSP, which leaves 131 usable features. Features containing references to image data are then filtered out, and columns containing string-type data with more than 5 unique values are filtered out.

To deal with missing entries in survey data, several imputation strategies are applied. For columns with numerical types of data, entries that are NaN will be replaced with the mean value, and then will have the L_2 norm applied to that column. Date/time types will have the median time imputed, and will then be scaled so the minimum and maximum are 0 and 1 respectively. String types—which we are treating as categorical types given the previous filtering of unique values—will be imputed with the most common value, and then split up into one-hot columns. With these steps, this gives 162 features upon which to train a model.

Various feature selection and classifier models were tested using the scikit-learn toolkit [14]. Classsifier models tested include XGBoost [31], linear support vector machine (SVM), multilayer perceptron, Gaussian NB, and k-nearest neighbors (k-NN). In addition to testing the proposed variations of mRMR, `SelectFromModel` (which selects based on the weights of a trained model) with linear SVM and XGBoost are tested. The linear SVM model is tested with C over 10 logarithmically spaced values from $[10^{-2}, 1]$, while the XGBoost `SelectFromModel` has the default settings.

An 80%/20% train/test split is used for evaluating overall performance, and 5-fold flat cross-validation is employed to both select hyperparameters and evaluate the overall performance of the dataset. Nested cross-validation is typically employed for evaluating model selection with small datasets, but experimentally may not be necessary with low numbers of hyper-parameters and using specific models, like gradient boosted trees [32]. For this reason, and due to the fact that nested cross-validation with K outer-folds would incur a K-fold increase in run-time, flat 5-fold cross-validation is used.

2.3 mRMR with mutual information (mRMR-i)

For the baseline implementation of mRMR, which shall be referred to as mRMR-i, the fast-mRMR implementation is used [13]. MI between features is calculated using the scikit-learn `mutual_info_regression` function, while MI between feature and label is calculated using `mutual_info_classif` [14], which are calculated using the methods described in section 1.2. The scikit-learn `fit/transform` design paradigm is followed for the implementation, and so the feature selection will occur within the `fit` stage and will be applied through calls to `transform`. The baseline mRMR-i is outlined in algorithm 1.

Algorithm 1 Fit the mRMR to X, y for a desired number of features N

```
procedure FIT($X, y$)
    selectedFeatures $\leftarrow \varnothing$
    candidates $\leftarrow 0...n_X$
    candidatesVec $\leftarrow \top : \forall x \in X$
    accumulatedRedundancy $\leftarrow 0 : \forall x \in X$

    relevancesVector $\leftarrow MI(X, y)$
    selected $\leftarrow \arg\max$ relevancesVector
    lastFeatureSelected $\leftarrow$ selected
    selectedFeatures $\leftarrow$ selectedFeatures $\cup$ selected
    candidates $\leftarrow$ candidates $\setminus$ selected

    while |selectedFeatures| $< N$ do
        max_mrmr $\leftarrow -\infty$
        newLastFeatureSelected $\leftarrow \varnothing$
        lastFeatureSelectedMI $\leftarrow MI(X_{candidatesVec}, X_{lastFeatureSelected})$

        for idxc $\leftarrow 1$ to $n$, can $\in$ candidatesVec do
            relevance $\leftarrow$ relevancesVector$_{can}$
            accumulatedRedundancy$_{can}$ $\leftarrow$ accumulatedRedundancy$_{can}$+ lastSelectedFeatureMI$_{idxc}$
            redundancy $\leftarrow$ accumulatedRedundancy$_{can}$/|selectedFeatures| mrmr $\leftarrow$ relevance $-$ redundancy

            if mrmr $>$ max_mrmr then
                max_mrmr $\leftarrow$ mrmr
                newLastFeatureSelected $\leftarrow$ can
            end if

            selectedFeatures $\leftarrow$ selectedFeatures $\cup$ newLastFeatureSelected
            candidates $\leftarrow$ candidates $\setminus$ newLastFeatureSelected
            candidatesVec$_{newLastFeatureSelected}$ $\leftarrow \perp$
            lastFeatureSelected $\leftarrow$ newLastFeatureSelected
        end for
    end while
end procedure
```

2.4 mRMR-s and mRMR-h

This section introduces two variations on the mRMR algorithm. mRMR with semantic textual similarity (mRMR-s) will utilize STS scores generated from a ClinicalBERT model as a direct replacement for MI scores. mRMR with hybrid mutual information and semantic textual similarity (mRMR-h) combines these two sources of information in assigning scores.

Fine-tuning on ClinicalBERT for STS would require a dataset with ground-truth labels for similar sentences. Two datasets, ClinicalSTS and MedSTS, are candidates for the fine-tuning [33], [34], but due to the logistical challenges in obtaining these datasets, we instead use the weights from a ClinicalBERT model fine-tuned on MedSTS [35]. The MedSTS dataset contains pairs with target labels defined on a 0 to 5 scale, with 5 representing identical meaning and 0 representing no shared semantic meaning. It is expected that most results will fall between 0 and 1, as a score of 1 is defined as "The two sentences are not equivalent, but are on the same topic" [34].

For mRMR with semantic textual similarity (mRMR-s), mutual information is replaced with the STS scores between feature questions and target questions computed using the aforementioned ClinicalBERT model. The resulting relevance and redundancy functions then become equations 2.1 and 2.2, while the underlying incremental search algorithm remains the same.

$$D = \frac{1}{|S|} \sum_{x_i \in S} \cos(E_i, E_c) \tag{2.1}$$

$$R = \frac{1}{|S|^2} \sum_{x_i, x_j \in S} \cos(E_i, E_j) \tag{2.2}$$

To combine STS and MI, mRMR-h involves treating the two scores as a linear combination, resulting in the equations for relevance and redundancy in equations 2.3 and 2.4. Only one hyperparameter is needed since there are two terms in the linear combination of MI and STS.

$$D = \frac{1}{|S|} \sum_{x_i \in S} I(x_i; c) + \alpha \cos(E_i, E_c) \tag{2.3}$$

$$R = \frac{1}{|S|^2} \sum_{x_i, x_j \in S} I(x_i, x_j) + \alpha \cos(E_i, E_j) \tag{2.4}$$

For evaluating the best hyperparameters in 5-fold cross-validation, we will use a linear space between $[0, 2]$ with 50 evenly-spaced values for the selection of α. To minimize the overall runtime of cross-validation, each fold is split across all available processors. Spawning a new process requires reloading a BERT model into the state of the program, and incurs a significant performance penalty each time. To avoid this overhead with each spawn, the STS scores among feature questions and between feature questions and target questions are pre-calculated and cached.

Chapter 3

Results

This chapter is begun by evaluating the overall performance of each possible model and feature selection configuration. After this, we look at the best possible hyperparameters for each model given the dataset. Finally, we examine the performance of the model on the given dataset as the hyperparmaeters are varied.

3.1 Model Performance

Table 3.1 highlights the performance of each combination of feature selector and model in terms of area under the receiver-operator curve (AUROC), area under the precision-recall curve (AUPRC), and accuracy. All feature selectors are fixed to select a total of 40 features, or approximately one-quarter of the possible features, for the purpose of comparison.

In terms of performance, we find that mRMR-s is the most effective variant of for XGBoost, while mRMR-h is the most effective for linear SVM. mRMR-h is a close second for Gaussian NB in terms of AUROC and AUPRC performance. mRMR-s and mRMR-h are the most effective mRMR variants for multilayer perceptron (MLP), while no variant of mRMR is effective for feature selection with k-NN.

3.1.1 Selected Features

Tables A.1, A.2, and A.3 contain the selected features from mRMR-i, mRMR-s, and mRMR-h, respectively. Each of the aformentioned tables shows the order in which the feature was selected, the assigned mRMR score as well as its constituent relevancy and redundancy scores. In addition, the feature importance when used in a Gaussian NB model is calculated

		AUROC		AUPRC	
Selector	Cls.	Test	Train	Test	Train
Identity	XGB	0.74253	0.852927	0.173495	0.267482
SFM-XGB	XGB	0.737628	0.855092	0.165651	0.247675
SFM-SVM	XGB	0.664566	0.862356	0.120219	**0.283392**
mRMR-i	XGB	0.689309	**0.890087**	0.124102	0.269925
mRMR-s	XGB	**0.814893**	0.875707	**0.20337**	0.255488
mRMR-h	XGB	0.744398	0.876307	0.1459	0.229443
Identity	SVM	0.898382	**0.920399**	**0.376203**	**0.449496**
SFM-XGB	SVM	0.87535	0.887429	0.283943	0.2783
SFM-SVM	SVM	0.875506	0.905066	0.226138	0.281013
mRMR-i	SVM	0.895736	0.913156	0.318761	0.361423
mRMR-s	SVM	0.897915	0.89419	0.334992	0.340455
mRMR-h	SVM	**0.908652**	0.906614	0.330394	0.354387
Identity	MLP	0.884532	0.935978	0.282288	0.516197
SFM-XGB	MLP	0.882353	0.881933	0.293784	0.277231
SFM-SVM	MLP	0.864768	0.901766	0.220033	0.277267
mRMR-i	MLP	0.880797	0.933514	0.27275	0.469809
mRMR-s	MLP	**0.888422**	0.897094	**0.310206**	0.348706
mRMR-h	MLP	0.882975	**0.938727**	0.272302	**0.513593**
Identity	GNB	0.80789	0.836726	0.157917	0.159329
SFM-XGB	GNB	**0.889823**	0.852209	0.279041	0.192056
SFM-SVM	GNB	0.824074	0.86878	0.177902	0.192038
mRMR-i	GNB	0.806178	0.843938	0.157338	0.166143
mRMR-s	GNB	0.868114	0.880422	**0.248683**	**0.238535**
mRMR-h	GNB	0.858077	**0.883347**	0.214441	0.229685
Identity	k-NN	0.823529	0.938004	0.2139	0.412918
SFM-XGB	k-NN	0.814581	0.93675	0.188433	0.38525
SFM-SVM	k-NN	**0.850373**	**0.940114**	**0.264685**	**0.431081**
mRMR-i	k-NN	0.713897	0.927701	0.151576	0.369271
mRMR-s	k-NN	0.778245	0.925156	0.195084	0.352407
mRMR-h	k-NN	0.747432	0.928896	0.15026	0.350174

Table 3.1: Results from using selected feature selection methods, all feature selection methods set to select up to 40 features. Best metrics for each model type and among mRMR are bolded.

using SHapley Additive exPlanations (SHAP) [36], and is shown in the last column. It should be noted that scores are not necessarily comparable between mRMR variants due to differences in scales between the scoring methods.

The underlying scores that are derived for feature-feature pairs and feature-target pairs using MI and STS, are shown in Figures 3.1 and 3.2, respectively.

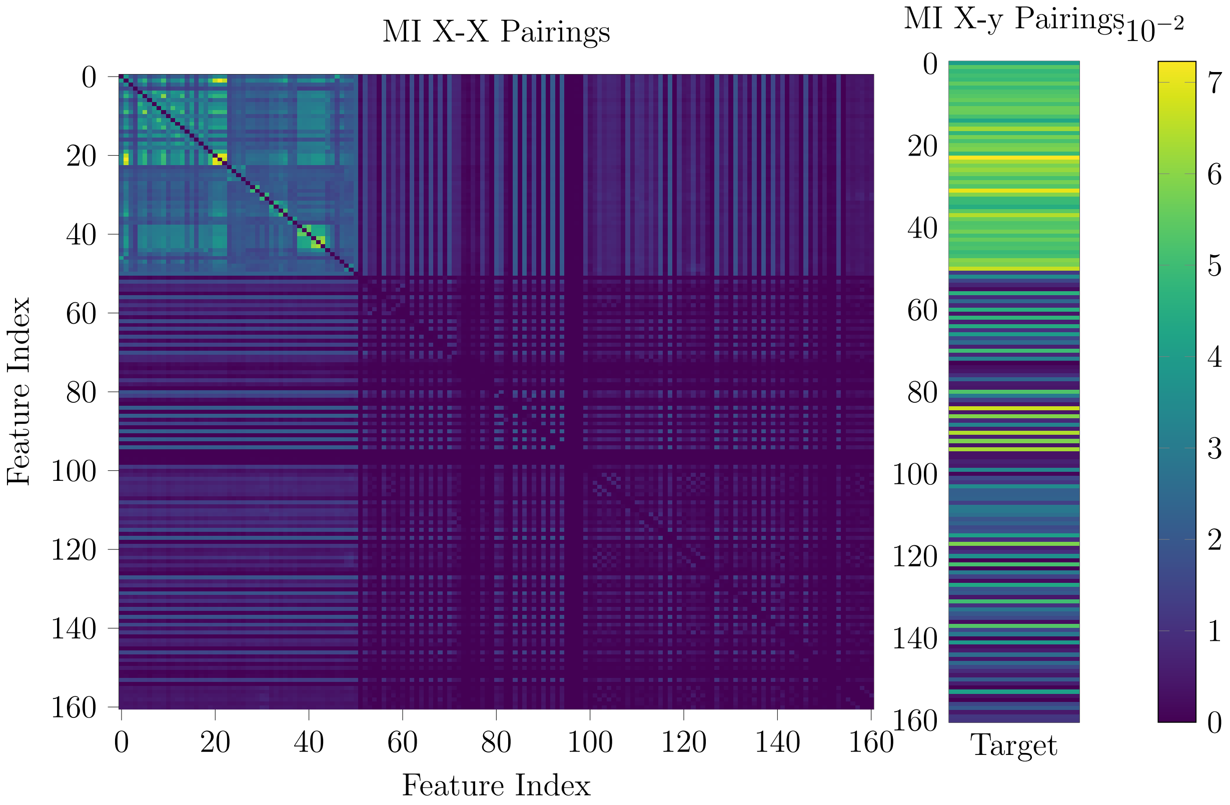


Figure 3.1: Heatmap of feature-feature and feature-target MI scores.

3.2 Hyperparameter Performance

3.2.1 Performance Over N For mRMR-i and mRMR-s

For mRMR-i and mRMR-s, there is only one hyperparameter, N, which is the number of features to select. Figures 3.3 and 3.5 show the train/test performance of Gaussian NB using the mRMR-i and mRMR-s feature selectors in terms of AUROC and AUPRC. In addition, figures 3.4 and 3.6 show the corresponding train/test performance using XGBoost with the same feature selectors.

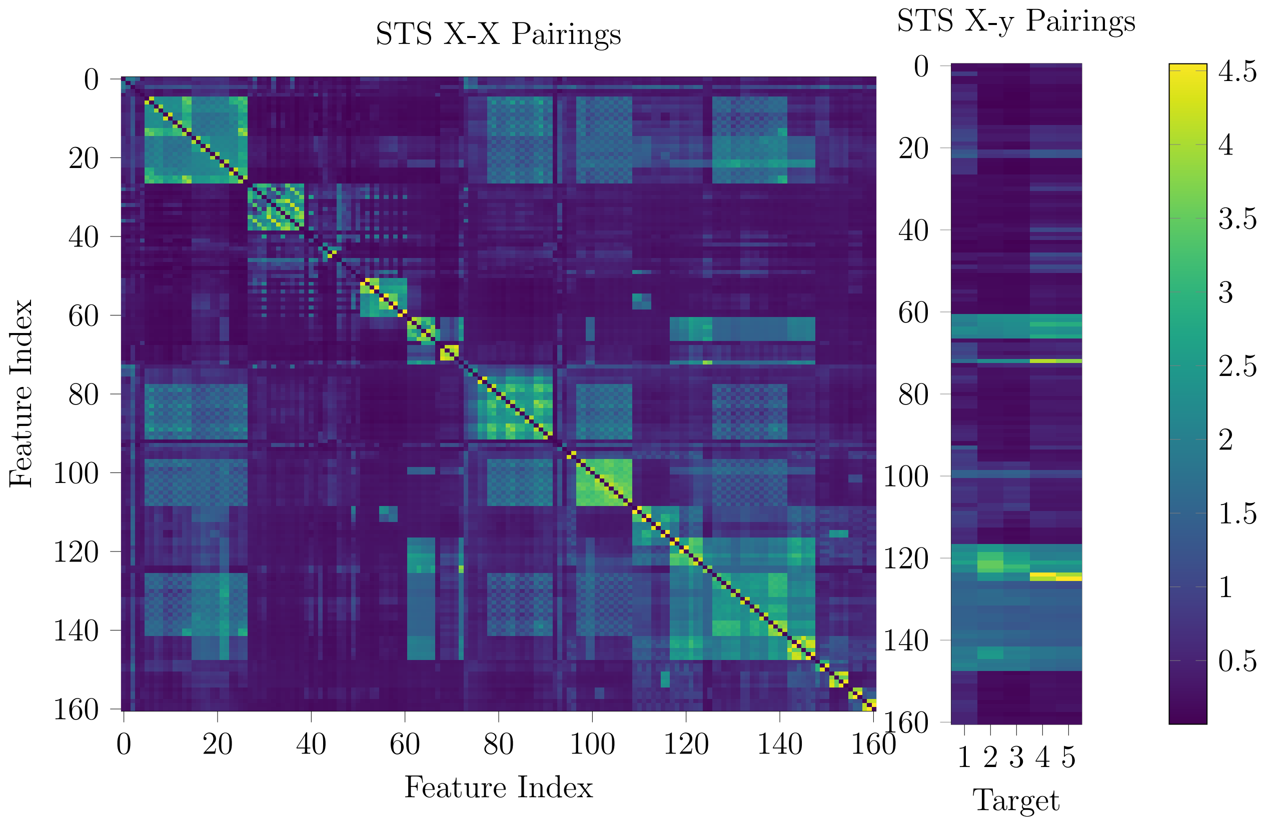
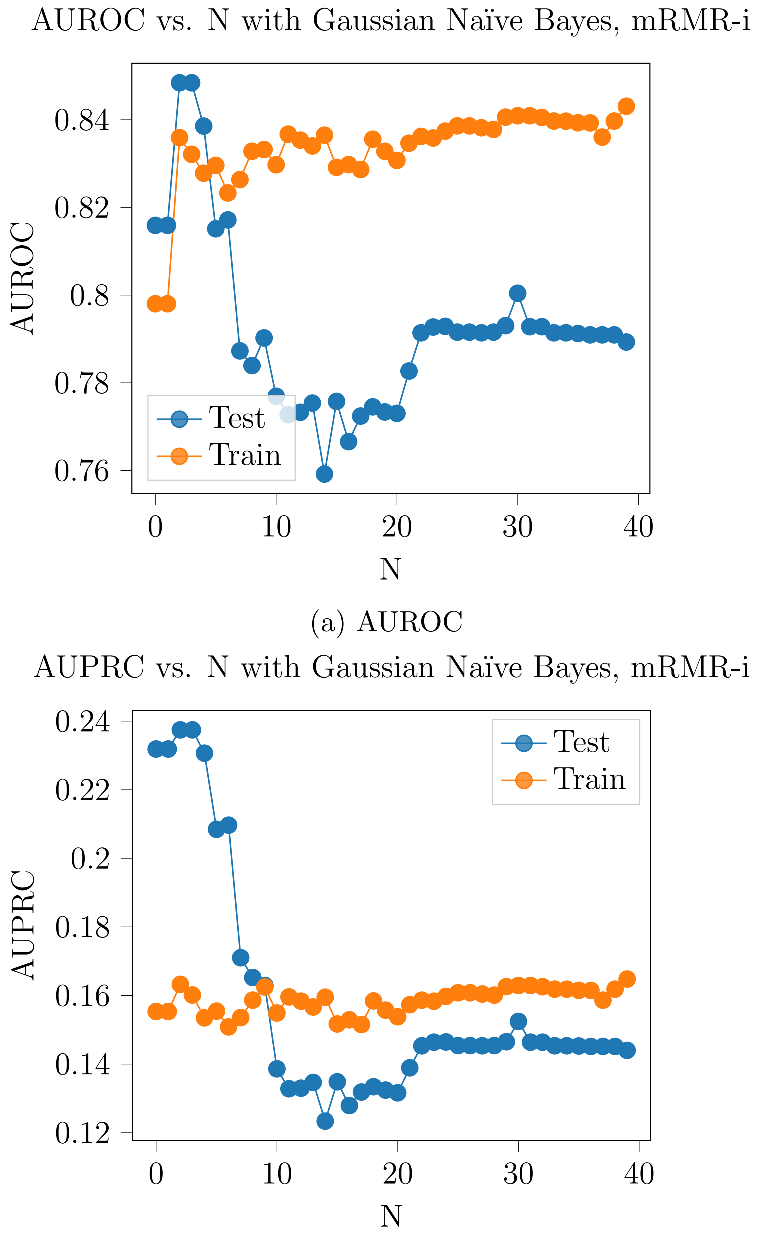


Figure 3.2: Heatmap of feature-feature and feature-target STS scores.

3.2.2 Performance Over N, α For mRMR-h

The linear combination of MI and STS in the corresponding R and D objective functions introduces another hyperparameter, α. Since there are only two elements of the linear combination, we leave MI unscaled and scale STS to remove a dimension from the hyperparameter space. To evaluate the performance, combinations of N and α were selected from the range $[0, 40)$ and 20 logaritmically-spaced numbers $[10^{-2}, 10^{1}]$, respectively. Figure 3.7 highlight the AUROC performance for a Gaussian NB classifier.

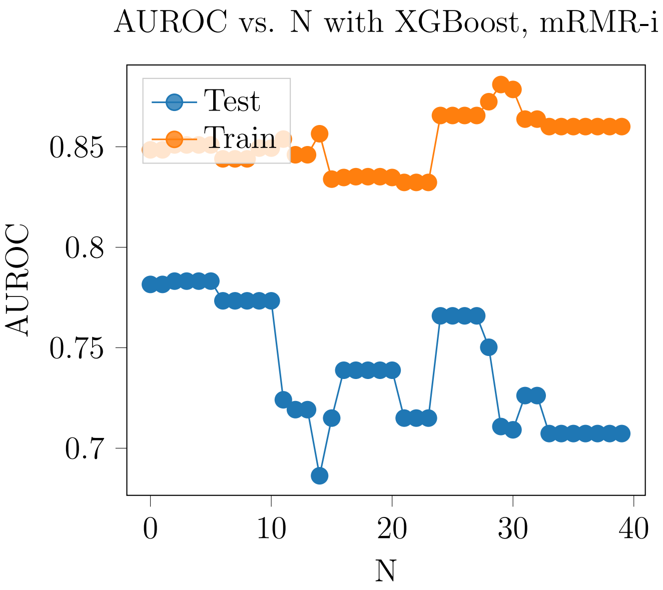
There are several noteworthy regions in Figure 3.7. The first noticeable region is the one between $\alpha \in [0.00, 0.41)$, where a noticeable improvement in AUROC occurs while requiring the utilization of a smaller number of questions. This suggests that some values of α result in MI-STS ratios that are more effective than others for selecting features to use.



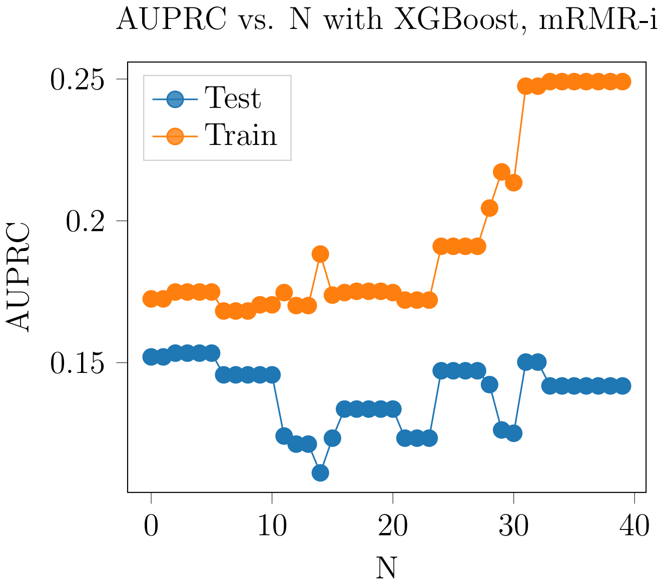
(a) AUROC

(b) AUPRC

Figure 3.3: Performance of a Gaussian NB classifier over N using mRMR-i

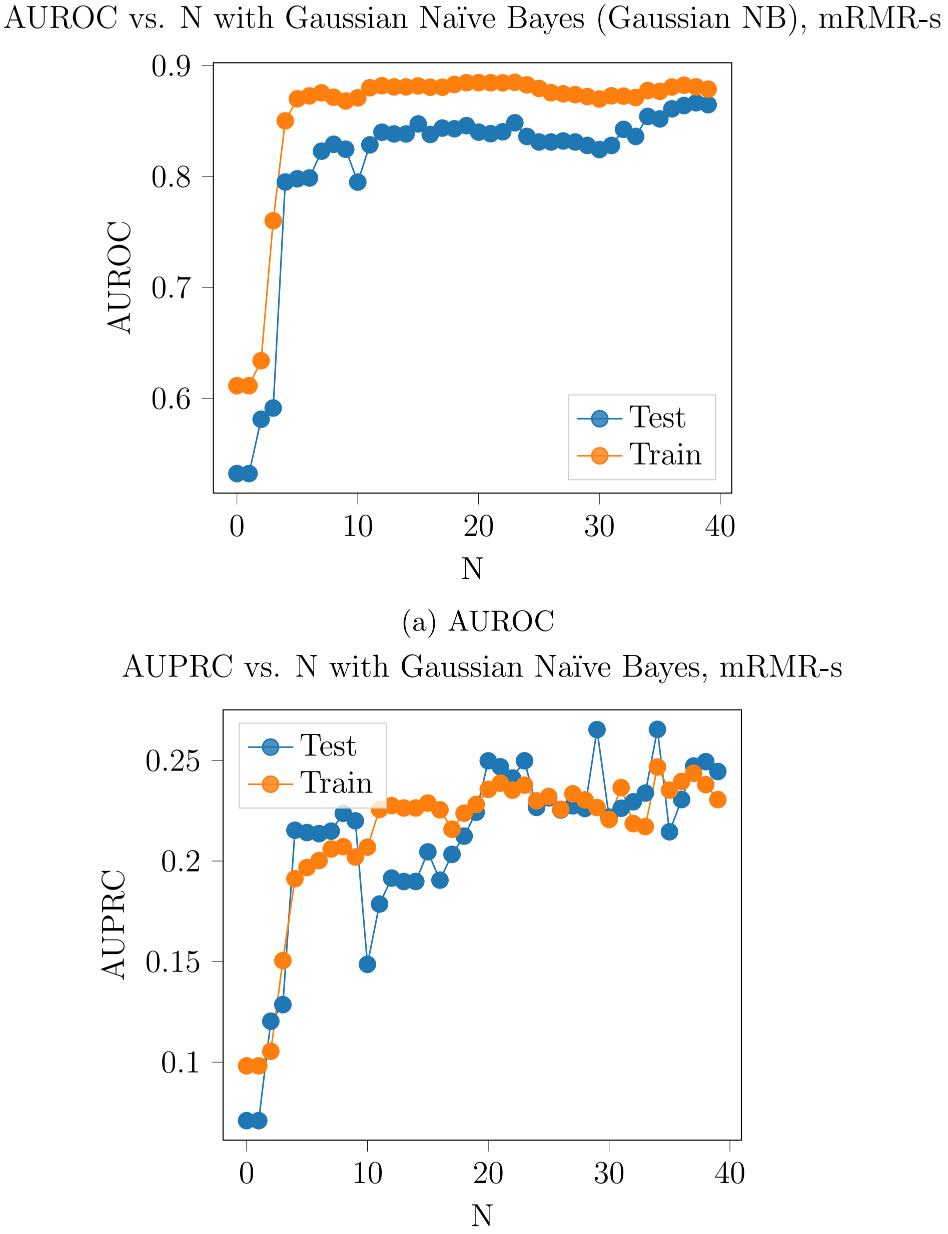


(a) AUROC



(b) AUPRC

Figure 3.4: Performance of a Gaussian NB classifier over N using mRMR-i

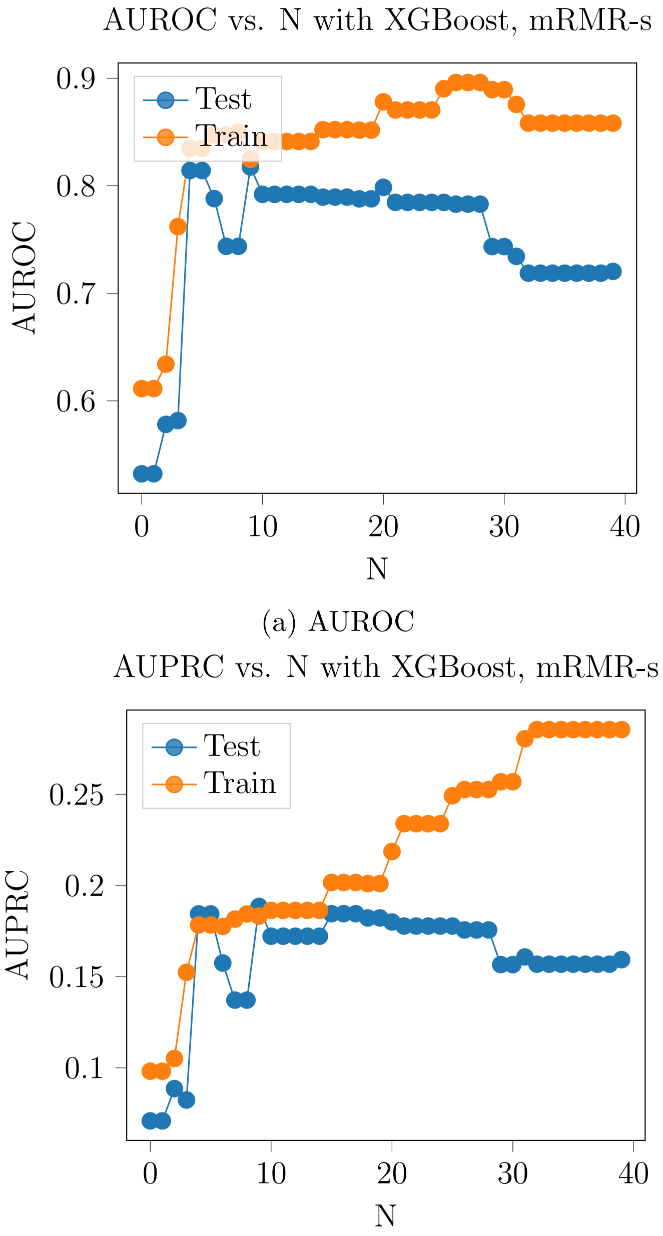

AUROC vs. N with Gaussian Naïve Bayes (Gaussian NB), mRMR-s

(a) AUROC

AUPRC vs. N with Gaussian Naïve Bayes, mRMR-s

(b) AUPRC

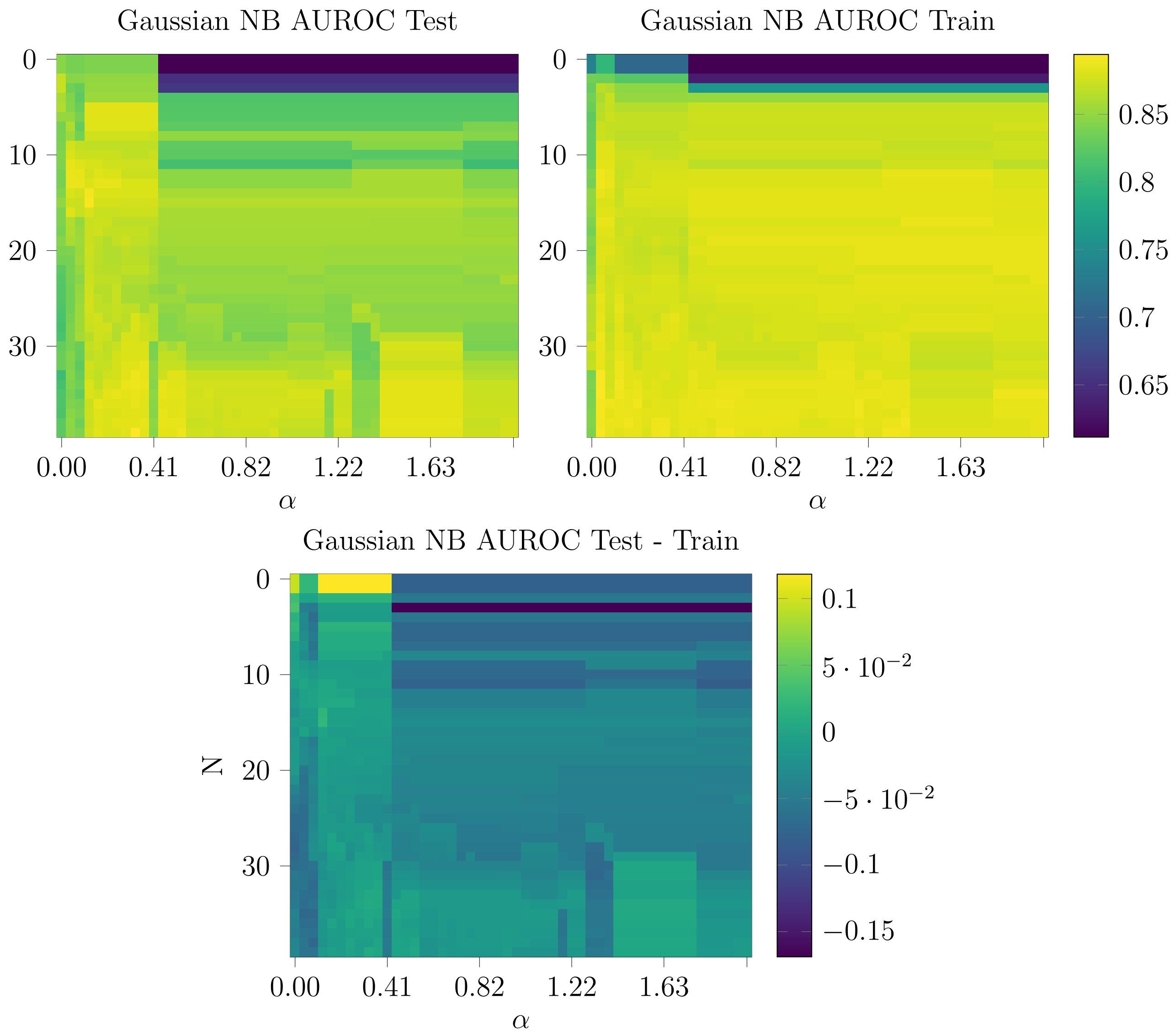
Figure 3.5: Performance of an XGBoost classifier over N using mRMR-s



(a) AUROC

(b) AUPRC

Figure 3.6: Performance of an XGBoost classifier over N using mRMR-s


Gaussian NB AUROC Test

Gaussian NB AUROC Train

Gaussian NB AUROC Test - Train

Figure 3.7: Test and train AUROC performance over the hyperparameter space for mRMR-h using Gaussian NB, as well as the difference between them.

Chapter 4

Discussion

This thesis is concluded with the discussion of model performance, as well as future opportunities for this area of work.

4.1 Model Performance

As briefly discussed in chapter 3, mRMR-s and mRMR-h appear to result in higher performance models and there are several possible reasons for this.

One principal reason that the usage of STS appears to work better than MI is that MI is only able to evaluate the relevance and redundancies between features one-on-one statistically. With many possible causes for the target label, individual features can share little MI, and features that we would expect to be more relevant than others will only have slightly more MI than those that would not be relevant. This becomes particularly evident for mRMR-i scores that dip into the negative: the subsequent feature learned is more redundant than it is relevant, yet there are many unselected features that we would consider to be truly relevant.

Features that end up having no effect in the end—as measured by SHAP value—still tend to share some MI with the target variable, as features with enough member examples will have some MI with the target variable. Some of the most striking examples of this can be seen in Table A.1. For example, there are four rows at the start with a measurably high amount of relevance, and no redundancy in relation to other features, that all end up having a SHAP value of 0 since the Gaussian NB model has picked up no true relationship between them and the target.

STS appears to be useful for several reasons. One reason is that STS is not always correlated with MI, meaning that it represents a contrasting, non-correlated, source of information compared to MI. This fact can be seen from the strong contrast in highlighted regions between Figures 3.1 and 3.2. STS is clearly more able to highlight redundant regions, especially along the diagonal, and is more able to distinguish between relevant and irrelevant features than MI. This is particularly useful for when MI that results from the training dataset fails to match what we might expect from the population sampled, as STS is not vulnerable to differences in the sample and population distributions.

Another reason why mRMR-s and mRMR-h appear to perform better is integration of clinical knowledge the selection of features. LLMs trained on clinical knowledge have demonstrated the ability to reason through question-answering problems [18], and the results here suggest that ClinicalBERT is able to connect feature concepts to target concepts. Of particular note are the features in Table A.2 that have a relevance score below 1. As discussed in section 2.4, a score of 1 is defined as "The two sentences are not equivalent, but are on the same topic" [34], meaning that values between 0 and 1 are weakly relevant. ClinicalBERT is able to ascertain that some features are relevant, such as "Age," "Final T-Score," and so on. Many of these variables do not appear among the features selected using mRMR-i, as seen in Table A.1, suggesting that STS is a useful source of information when metric like MI cannot fully capture the relationships underpinning a dataset.

4.2 Future Work

One area of future work could consider the mRMR algorithm using metrics other than MI or STS, as both metrics have weaksnesses that make them ineffective metrics. MI is incapable of measuring the true amount of information a feature contains in context with other features, and STS can only be used to represent semantic relationships rather than true statistical relationships.

Future work could also consider the choice of model for computing STS. Many new pre-trained transformer models have been released since the original ClinicalBERT model was released in 2019 [25], such as BioGPT [37], PubMedBERT [38], and GatorTron [39], and further work could explore how these different architectures perform when used in mRMR-s and mRMR-h.

Future work could also consider the use of semantic pairs that specifically rate *relevancy* and *redundancy* between pairs of question embeddings, rather than similarity. Relevant and redundant questions may not always be semantically similar, and a dataset for fine-tuning upon this type of task may improve the performance of mRMR-s or mRMR-h.

Another potential area of future work would be to serialize the mRMR objective into a text prompt. Serialization of tablular data into a question prompt for a large language model can achieve high performance in a few-shot learning context [40], and serialization of the mRMR objective may also be able to capture further semantic relationships between features.

References

[1] M. R. Vila, M. S. Todorovic, C. Tang, *et al.*, “Cognitive flexibility and persistent post-surgical pain: The flexcapp prospective observational study,” *British journal of anaesthesia*, vol. 124, no. 5, pp. 614–622, 2020.

[2] S. Haroutiunian, L. Nikolajsen, N. B. Finnerup, and T. S. Jensen, “The neuropathic component in persistent postsurgical pain: A systematic literature review,” *PAIN®*, vol. 154, no. 1, pp. 95–102, 2013.

[3] P. A. Harris, R. Taylor, R. Thielke, J. Payne, N. Gonzalez, and J. G. Conde, “Research electronic data capture (redcap)—a metadata-driven methodology and workflow process for providing translational research informatics support,” *Journal of biomedical informatics*, vol. 42, no. 2, pp. 377–381, 2009.

[4] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, *Feature extraction: foundations and applications*. Springer, 2008, vol. 207.

[5] B. Remeseiro and V. Bolon-Canedo, “A review of feature selection methods in medical applications,” *Computers in biology and medicine*, vol. 112, p. 103 375, 2019.

[6] P. Washington, K. M. Paskov, H. Kalantarian, *et al.*, “Feature selection and dimension reduction of social autism data,” in *Pacific Symposium on Biocomputing 2020*, 2019, pp. 707–718.

[7] U. Niemann, P. Brueggemann, B. Boecking, B. Mazurek, and M. Spiliopoulou, “Development and internal validation of a depression severity prediction model for tinnitus patients based on questionnaire responses and socio-demographics,” *Scientific reports*, vol. 10, no. 1, p. 4664, 2020.

[8] H. Abbas, F. Garberson, E. Glover, and D. P. Wall, “Machine learning approach for early detection of autism by combining questionnaire and home video screening,” *Journal of the American Medical Informatics Association*, vol. 25, no. 8, pp. 1000–1007, 2018.

[9] V. S. Saridewi and R. F. Sari, “Feature selection in the human aspect of information security questionnaires using multicluster feature selection,” *International Journal of Advanced Science and Technology*, vol. 29, no. 7 Special Issue, pp. 3484–3493, 2020.

[10] Y. Chen, K. Wang, and J. J. Lu, “Feature selection for driving style and skill clustering using naturalistic driving data and driving behavior questionnaire,” *Accident Analysis & Prevention*, vol. 185, p. 107 022, 2023.

[11] F. Abut, M. F. Akay, and J. George, “Developing new vo2max prediction models from maximal, submaximal and questionnaire variables using support vector machines combined with feature selection,” *Computers in biology and medicine*, vol. 79, pp. 182–192, 2016.

[12] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[13] S. Ramírez-Gallego, I. Lastra, D. Martínez-Rego, *et al.*, "Fast-mrmr: Fast minimum redundancy maximum relevance algorithm for high-dimensional big data," *International Journal of Intelligent Systems*, vol. 32, no. 2, pp. 134–152, 2017.

[14] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[15] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical review E*, vol. 69, no. 6, p. 066 138, 2004.

[16] B. C. Ross, "Mutual information between discrete and continuous data sets," *PloS one*, vol. 9, no. 2, e87357, 2014.

[17] C. D. Manning, "Human language understanding & reasoning," *Daedalus*, vol. 151, no. 2, pp. 127–138, 2022.

[18] K. Singhal, S. Azizi, T. Tu, *et al.*, "Large language models encode clinical knowledge," *arXiv preprint arXiv:2212.13138*, 2022.

[19] C. Wei, Y.-C. Wang, B. Wang, and C.-C. J. Kuo, "An overview on language models: Recent developments and outlook," *arXiv preprint arXiv:2303.05759*, 2023.

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[21] D. Erhan, A. Courville, Y. Bengio, and P. Vincent, "Why does unsupervised pre-training help deep learning?" In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, JMLR Workshop and Conference Proceedings, 2010, pp. 201–208.

[22] T. Wolf, L. Debut, V. Sanh, *et al.*, "Huggingface's transformers: State-of-the-art natural language processing," *arXiv preprint arXiv:1910.03771*, 2019.

[23] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[24] D. Oniani, S. Sivarajkumar, and Y. Wang, "Few-shot learning for clinical natural language processing using siamese neural networks," *arXiv preprint arXiv:2208.14923*, 2022.

[25] E. Alsentzer, J. R. Murphy, W. Boag, *et al.*, "Publicly available clinical bert embeddings," *arXiv preprint arXiv:1904.03323*, 2019.

[26] A. E. Johnson, T. J. Pollard, L. Shen, *et al.*, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[27] E. Lehman, E. Hernandez, D. Mahajan, *et al.*, "Do we still need clinical language models?" *arXiv preprint arXiv:2302.08091*, 2023.

[28] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[29] V. Lampos, B. Zou, and I. J. Cox, "Enhancing feature selection using word embeddings: The case of flu surveillance," in *Proceedings of the 26th International Conference on World Wide Web*, 2017, pp. 695–704.

[30] A. Roy and S. Pan, "Incorporating medical knowledge in bert for clinical relation extraction," in *Proceedings of the 2021 conference on empirical methods in natural language processing*, 2021, pp. 5357–5366.

[31] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[32] J. Wainer and G. Cawley, "Nested cross-validation when selecting classifiers is overzealous for most practical applications," *Expert Systems with Applications*, vol. 182, p. 115 222, 2021.

[33] Y. Wang, S. Fu, F. Shen, S. Henry, O. Uzuner, H. Liu, *et al.*, "The 2019 n2c2/ohnlp track on clinical semantic textual similarity: Overview," *JMIR medical informatics*, vol. 8, no. 11, e23375, 2020.

[34] Y. Wang, N. Afzal, S. Fu, *et al.*, "Medsts: A resource for clinical semantic textual similarity," *Language Resources and Evaluation*, vol. 54, pp. 57–72, 2020.

[35] A. Mulyar, E. Schumacher, and M. Dredze, *Semantic-text-similarity*, 2019. [Online]. Available: https://github.com/AndriyMulyar/semantic-text-similarity.

[36] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[37] R. Luo, L. Sun, Y. Xia, *et al.*, "Biogpt: Generative pre-trained transformer for biomedical text generation and mining," *Briefings in Bioinformatics*, vol. 23, no. 6, 2022.

[38] Y. Gu, R. Tinn, H. Cheng, *et al.*, "Domain-specific language model pretraining for biomedical natural language processing," *ACM Transactions on Computing for Healthcare (HEALTH)*, vol. 3, no. 1, pp. 1–23, 2021.

[39] X. Yang, A. Chen, N. PourNejatian, *et al.*, "A large language model for electronic health records," *npj Digital Medicine*, vol. 5, no. 1, p. 194, 2022.

[40] S. Hegselmann, A. Buendia, H. Lang, M. Agrawal, X. Jiang, and D. Sontag, "Tabllm: Few-shot classification of tabular data with large language models," *arXiv preprint arXiv:2210.10723*, 2022.

Appendix A

Feature Importance

Feature Name	n	mRMR Score	Relevancy	Redundancy	Mean Absolute SHAP
Current Medications: Confirmed with participant (choice=Metformin)_Checked	1	0.0722903	0.0722903	0	0
Please specify your race: (choice=Black / African Heritage)_Checked	2	0.0668161	0.0668161	0	0
Please specify your race: (choice=Hawaiian Native / Other Pacific Islander)_Checked	3	0.0584943	0.0584943	0	0
Please specify your race: (choice=American Indian / Alaskan Native)_Checked	4	0.0522559	0.0522559	0	0
Baseline Emotional Distress-Anxiety - Short Form 4a T Score	5	0.046146	0.046146	0	1.22624e-05
Baseline Cognitive Function - Abilities - Short Form 4a T Score	6	0.011949	0.0351969	0.0232478	0
Please specify your ethnicity:_Non-Hispanic	7	-0.00107861	0.0227256	0.0238042	0.0118578

Did you find submerging your hand in the cold water bath to be:_Not Painful	8	-0.00486099	0.0496789	0.0545398	0
Pain intensity - Baseline 2:	9	-0.0145096	0.0474391	0.0619487	0
Pain intensity - CPM Average:	10	-0.011965	0.0400014	0.0519664	0
Baseline Emotional Distress-Depression - Short Form 4a T Score	11	-0.0159725	0.0240158	0.0399883	0
How often have you been using cannabis (marijuana, pot, weed, grass) in the past year?_Never	12	-0.0286325	0.0408129	0.0694454	0.00290619
Pain intensity - CPM 1:	13	-0.0478522	0.0445602	0.0924124	0
Moderate activities such as moving a table, pushing a vacuum cleaner, bowling, or playing golf?_Yes limited a lot	14	-0.0585221	0.0203294	0.0788515	0.000588596
Climbing several flights of stairs?_Yes limited a lot	15	-0.0591006	0.0109577	0.0700583	0.000269773
Why are you having this upcoming procedure? Please mark all that apply. Please mark all that apply. (choice=My doctor said I needed the procedure)_Unchecked	16	-0.0536297	0.0292872	0.082917	0.000637646
Do you take any of the following medications for pain treatment at least once per week? (choice=Tramadol (Ultram))_Unchecked	17	-0.051138	0.0501994	0.101337	0.00564071

Did you find submerging your hand in the cold water bath to be:_Mildly Painful	18	-0.0523673	0.0249529	0.0773202	0
Have you ever been diagnosed with anxiety?_No	19	-0.059889	0.0288636	0.0887526	0
Have you ever been diagnosed with anxiety?_Yes	20	-0.0611281	0.0262699	0.087398	0.00174126
Do you take any of the following medications for pain treatment at least once per week? (choice=Duloxetine (Cymbalta) or amitriptyline (Elavil))_Unchecked	21	-0.0590742	0.0530448	0.112119	0.008645
Have you ever been diagnosed with chronic pain?_No	22	-0.0644443	0.0579427	0.122387	0
Moderate activities such as moving a table, pushing a vacuum cleaner, bowling, or playing golf?_Yes limited a little	23	-0.0619967	0.0151508	0.0771475	0.000821582
On a scale of zero to ten, with zero being no pain and ten being the worst pain, what is your average pain level at rest TODAY?	24	-0.0658826	0.031516	0.0973986	0
Please specify your race: (choice=Black / African Heritage)_Unchecked	25	-0.0672812	0.00402699	0.0713082	0.00459841
Please specify your race: (choice=Prefer not to answer)_Checked	26	-0.0628143	0.0638041	0.126618	0

Please specify your race: (choice=Other)_Checked	27	-0.0617346	0.0325781	0.0943127	0
Baseline Physical Function - Short Form 4a	28	-0.0646048	0.0151008	0.0797057	0
Do you take any of the following medications for pain treatment at least once per week? (choice=Aspirin)_Unchecked	29	-0.0648413	0.0430928	0.107934	0.00851012
Have you been experiencing pain in the past week?_Yes, and RELATED to my need for surgery	30	-0.0648443	0.0513111	0.116155	0
Total Years of Education	31	-0.0662292	0.0581167	0.124346	0
Have you ever been diagnosed with a Post Traumatic Stress Disorder (PTSD)?_Yes	32	-0.0701365	0.0182878	0.0884243	0
Other:.1_married; polyam	33	-0.0696431	0.062555	0.132198	0
In the past three months, did you experience daily or near daily pain?_Yes	34	-0.0687923	0.0362142	0.105007	0.00166769
Do you take any of the following medications for pain treatment at least once per week? (choice=Gabapentin or Pregabalin)_Unchecked	35	-0.0764872	0.0161022	0.0925894	0.00229307
Have you ever been diagnosed with depression?_Yes	36	-0.0754841	0.0226042	0.0980883	0.00121398
Please specify your race: (choice=Caucasian)_Unchecked	37	-0.076869	0.00424245	0.0811114	0
Please specify your ethnicity:_Hispanic	38	-0.0746797	0.00940563	0.0840853	0.000735745

Do you take any of the following medications for pain treatment at least once per week? (choice=None)_Unchecked	39	-0.0733169	0.0390622	0.112379	0.00349479
What is your sex (assigned at birth):_Female	40	-0.0740775	0.00283702	0.0769145	0.00268547

Table A.1: Feature importance using Gaussian NB for mRMR-i with 40 features

Feature Name	n	mRMR	Relevancy	Redundancy	Mean Absolute SHAP
On a scale of zero to ten, with zero being no pain and ten being the worst pain , please mark your average pain level during the past week, while you were active or moving.	1	2.92961	2.92961	0	0.00166769
Do you take any of the following medications for pain treatment at least once per week? (choice=Aspirin)_Checked	2	1.32006	1.32006	0	0.00185162
Do you take any of the following medications for pain treatment at least once per week? (choice=Acetaminophen (Tylenol))_Checked	3	0.717374	1.49348	0.776111	0.00307787

Have you been experiencing pain in the past week?_Yes, and RELATED to my need for surgery	4	0.633957	2.65388	2.01992	0.00264868
Have you been experiencing pain in the past week?_Yes, but NOT RELATED to my need for surgery	5	1.02072	2.54984	1.52912	0.0029675
On a scale of zero to ten, with zero being no pain and ten being the worst pain, what is your average pain level at rest TODAY?	6	0.415542	2.80485	2.38931	0
Pain intensity - CPM 1:	7	0.161247	2.26003	2.09878	0
Pain intensity - CPM 2:	8	0.440164	2.2078	1.76763	0
In the past three months, did you experience daily or near daily pain?_No	9	0.355732	2.3925	2.03676	0.00112814
Have you ever been diagnosed with chronic pain?_No	10	0.468039	2.06852	1.60049	0.00339669
In the past three months, did you experience daily or near daily pain?_Yes	11	0.362483	2.39647	2.03399	0.00472103
On a scale of zero to ten, with zero being no pain and ten being the worst pain, please mark your average pain level during the past week, while at rest.	12	0.492365	2.89288	2.40052	0.00167995
Have you been experiencing pain in the past week?_No	13	0.59769	2.80017	2.20248	0.00134887

Have you ever been diagnosed with chronic pain?_Yes	14	0.459705	2.06441	1.6047	0.00179031
Why are you having this upcoming procedure? Please mark all that apply. Please mark all that apply. (choice=Decrease pain)_Checked	15	0.200513	1.27984	1.07933	0
Why are you having this upcoming procedure? Please mark all that apply. Please mark all that apply. (choice=Decrease pain)_Unchecked	16	0.180732	1.18812	1.00739	0
Pain intensity - Baseline Average:	17	0.211212	2.42521	2.214	0.00126303
Have you used opioid medications for pain management in the past?_yes	18	0.146445	1.96976	1.82331	0.00274678
Pain intensity - CPM Average:	19	0.106595	2.31014	2.20355	0
Final T-Score	20	0.11073	0.562743	0.452013	1.22624e-05
Total Score	21	0.108315	0.661823	0.553508	0
Have you used opioid medications for pain management in the past?_no	22	0.0913118	1.96299	1.87168	0.00207235
Pain intensity - Baseline 2:	23	0.180321	2.27706	2.09674	0.00137339
Pain intensity - Baseline 1:	24	0.227999	2.33905	2.11105	0
Interference	25	0.0677081	0.328047	0.260339	0
Age:	26	0.0691869	0.366818	0.297631	0.00174126

Do you take any of the following medications for pain treatment at least once per week? (choice=Aspirin)_Unchecked	27	0.103845	1.2731	1.16926	0.0184059
Have you used opioid medications for pain management in the past?_I don't know	28	0.0940281	1.94502	1.85099	0.000465972
Rumination Score	29	0.0938397	0.50304	0.4092	0
Do you currently use opioid medications for pain management?_no	30	0.104262	1.7738	1.66954	0.00605763
Helplessness Score	31	0.109871	0.694804	0.584933	0
Do you currently use opioid medications for pain management?_I don't know	32	0.144523	1.78068	1.63616	0.00076027
Do you take any of the following medications for pain treatment at least once per week? (choice=Tramadol (Ultram))_Unchecked	33	0.137729	1.53953	1.4018	0.018553
Do you currently use opioid medications for pain management?_yes	34	0.125729	1.73716	1.61143	0.00445126
BRS Score	35	0.107977	0.496375	0.388397	0
Do you take any of the following medications for pain treatment at least once per week? (choice=Gabapentin or Pregabalin)_Checked	36	0.111957	1.36527	1.25332	0.00282036

Do you take any of the following medications for pain treatment at least once per week? (choice=Tramadol (Ultram))_Checked	37	0.107958	1.53492	1.42696	0.00207235
Do you take any of the following medications for pain treatment at least once per week? (choice=Acetaminophen (Tylenol))_Unchecked	38	0.103917	1.48786	1.38394	0.00239117
Do you take any of the following medications for pain treatment at least once per week? (choice=Other)_Unchecked	39	0.0531368	1.53429	1.48115	0.00347026
Magnification Score	40	0.0485151	0.351817	0.303301	0

Table A.2: Feature importance using Gaussian NB using mRMR-s with 40 features.

Feature Name	n	mRMR	Relevancy	Redundancy	Mean Absolute SHAP
On a scale of zero to ten, with zero being no pain and ten being the worst pain , please mark your average pain level during the past week, while you were active or moving.	1	29.3107	29.3107	0	0.00237891

Do you take any of the following medications for pain treatment at least once per week? (choice=Aspirin)_Checked	2	12.9	13.2062	0.306242	0.00185162
Do you take any of the following medications for pain treatment at least once per week? (choice=Acetaminophen (Tylenol))_Checked	3	7.13422	14.9638	7.8296	0.00248927
Have you been experiencing pain in the past week?_Yes, and RELATED to my need for surgery	4	6.34058	26.5799	20.2393	0.00213366
Have you been experiencing pain in the past week?_Yes, but NOT RELATED to my need for surgery	5	10.1997	25.4984	15.2987	0.00284488
On a scale of zero to ten, with zero being no pain and ten being the worst pain, what is your average pain level at rest TODAY?	6	4.14844	28.0843	23.9359	0
Pain intensity - CPM 1:	7	1.55363	22.6376	21.084	0
Pain intensity - CPM 2:	8	4.37524	22.0825	17.7073	0
In the past three months, did you experience daily or near daily pain?_No	9	3.48939	23.9295	20.4401	0.000686695
Have you ever been diagnosed with chronic pain?_No	10	4.56223	20.7428	16.1806	0.00364194

In the past three months, did you experience daily or near daily pain?_Yes	11	3.60014	23.9924	20.3922	0.00478234
On a scale of zero to ten, with zero being no pain and ten being the worst pain, please mark your average pain level during the past week, while at rest.	12	4.85349	28.9413	24.0878	0
Have you been experiencing pain in the past week?_No	13	5.92479	28.0017	22.077	0.00196199
Have you ever been diagnosed with chronic pain?_Yes	14	4.59072	20.6562	16.0655	0.00197425
Why are you having this upcoming procedure? Please mark all that apply. Please mark all that apply. (choice=Decrease pain)_Checked	15	1.93161	12.8317	10.9001	0
Why are you having this upcoming procedure? Please mark all that apply. Please mark all that apply. (choice=Decrease pain)_Unchecked	16	1.78804	11.8824	10.0944	0
Pain intensity - Baseline Average:	17	2.08934	24.2521	22.1628	0.00139792
Have you used opioid medications for pain management in the past?_yes	18	1.43209	19.7206	18.2885	0.00313918
Pain intensity - CPM Average:	19	0.977467	23.1422	22.1647	0

Final T-Score	20	0.958883	5.67918	4.7203	0
Total Score	21	0.914266	6.66802	5.75375	0
Have you used opioid medications for pain management in the past?_no	22	0.917278	19.6458	18.7286	0.00324954
Pain intensity - Baseline 2:	23	1.67049	22.8134	21.1429	0.00240343
Pain intensity - Baseline 1:	24	2.26412	23.3986	21.1345	0
Interference	25	0.444967	3.32661	2.88164	0
Do you take any of the following medications for pain treatment at least once per week? (choice=Aspirin)_Unchecked	26	0.705992	12.78	12.074	0.0158676
Rumination Score	27	0.482619	5.08523	4.60261	0
Do you currently use opioid medications for pain management?_no	28	0.917341	17.7617	16.8444	0.00535868
Age:	29	0.536772	3.70914	3.17237	0.00107909
Have you used opioid medications for pain management in the past?_I don't know	30	1.06422	19.4564	18.3922	0.000429185
Do you take any of the following medications for pain treatment at least once per week? (choice=Tramadol (Ultram))_Unchecked	31	1.00569	15.438	14.4323	0.0183078
Do you currently use opioid medications for pain management?_I don't know	32	1.20055	17.8068	16.6063	0.00104231
Helplessness Score	33	1.0705	6.9977	5.9272	0

Do you currently use opioid medications for pain management?_yes	34	0.789919	17.3952	16.6053	0.0041447
Do you take any of the following medications for pain treatment at least once per week? (choice=Acetaminophen (Tylenol))_Unchecked	35	0.860962	14.8857	14.0248	0.00180258
Do you take any of the following medications for pain treatment at least once per week? (choice=Gabapentin or Pregabalin)_Checked	36	0.898638	13.6626	12.764	0.00282036
BRS Score	37	0.519309	5.02212	4.50281	0
Do you take any of the following medications for pain treatment at least once per week? (choice=Tramadol (Ultram))_Checked	38	0.675679	15.3492	14.6735	0.00207235
Do you take any of the following medications for pain treatment at least once per week? (choice=Ibuprofen (Motrin or Advil), celecoxib (Celebrex, Naproxen or Aleve), or other NSAIDs)_Unchecked	39	0.397347	13.6129	13.2156	0.00361741
Do you take any of the following medications for pain treatment at least once per week? (choice=Other)_Unchecked	40	0.355878	15.3664	15.0106	0.0032618

Table A.3: Feature importance using Gaussian NB using mRMR-h with 40 features. Features are selected with $\alpha = 0.0923671$, which was found through 5-fold flat cross-validation.