Utilizing Semantic Textual Similarity for Clinical Survey Data Feature Selection

Benjamin C. Warner, Ziqi Xu, Simon Haroutounian, Thomas Kannampallil, Chenyang Lu
Washington University in St. Louis
{b.c.warner,ziqixu,sharout,thomas.k,lu}@wustl.edu

Abstract

Surveys are widely used to collect patient data in healthcare, and there is significant clinical interest in predicting patient outcomes using survey data. However, surveys often include numerous features that lead to high-dimensional inputs for machine learning models. This paper exploits a unique source of information in surveys for feature selection. We observe that feature names (i.e., survey questions) are often semantically indicative of what features are most useful. Using language models, we leverage semantic textual similarity (STS) scores between features and targets to select features. The performance of STS scores in directly ranking features as well as in the minimal-redundancy-maximal-relevance (mRMR) algorithm is evaluated using survey data collected as part of a clinical study on persistent post-surgical pain (PPSP) as well as an accessible dataset collected through the NIH *All of Us* program. Our findings show that features selected with STS can result in higher performance models compared to traditional feature selection algorithms.

1 Introduction

Survey data can have many features and a relatively low number of examples, particularly in human subjects research where there is a high cost of recruiting participants. Small and novel populations combined with multiple externally-validated questionnaires can easily result in high-dimensional data. A typical solution is to apply feature selection using statistical measures, but because this relies upon the same high-dimensional data being filtered, suboptimal selections can be made. Conversely, features can be manually selected for inclusion in a model, but this can be limited as it involves testing individual combinations of features. An ideal solution would combine the empirical ranking of the former approach with the domain-informed selection offered by the latter approach.

We explore this problem through two clinical datasets, a private dataset from a study examining persistent post-surgical pain (PPSP) and the *All of Us* dataset. Both of these datasets collect answers to hundreds of questions filled out by clinicians and participants covering a broad range of topics. For the PPSP dataset, there are only 617 examples, which limits the amount of information available for feature selection.

1.1 Proposed Approach

Survey questions are, in practice, semantically related to a target outcome and semantically similar to or distinct from other questions. Semantic textual similarity (STS) could be an analogue to statistical measures that capture relationships between features, such as mutual information (MI). STS may then be useful for determining which questions are *relevant* to predicting a target question, or which questions are *redundant* to one another. To this end, we evaluate the use of STS scores directly to determine the most relevant features, and test STS scores both as a direct replacement and as a complement in algorithms utilizing statistical scores, such as minimal-redundancy-maximal-relevance (mRMR). Our overall approach, outlined in Figure 1, is to first score features using a language model (LM) that produces STS scores, statistical scores, or a linear combination of both; and then use a general selection algorithm to pick features from these scores.

There appears to be nearly no literature examining feature selection with embeddings or STS. The closest match to our knowledge examined the usage of word2vec continuous bag-of-words embeddings (Mikolov et al., 2013) trained upon Twitter data to select Google search query trends matching the embeddings of a target concept (Lampos et al., 2017). Our approach differs in several key ways, with the principal difference being that we use STS scores to make comparative selections with limited tab-

Findings of the Association for Computational Linguistics: ACL 2025, pages 502–520
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

ular data, whereas (Lampos et al., 2017) use the distribution of STS scores produced by thousands of candidate time-series features to make selections. In addition, we compare the performance of different selection algorithms, pre-trained LMs, fine-tuning datasets and scoring combinations as they relate to feature selection performance.

We present several empirical contributions:

- An examination of how STS scores generated by LMs between feature-feature and feature-target question pairs can help select features, either alone or in combination with statistical measures such as MI.
- A comparison of different LM scoring configurations as they relate to classifier performance.
- A demonstration of how STS-based feature selection algorithms can reduce overfitting using a gated-access and private clinical dataset relating to persistent post-surgical pain.

We make our code available at `https://github.com/bcwarner/sts-select` and via `pip install sts-select`. Additionally, a Hugging Face collection with the best performing models for each dataset can be found through our GitHub.

2 Preliminaries

2.1 Feature Selection

Fitting high-dimensional data is particularly difficult when the number of examples is low since a model can easily overfit on the training data. To counter this, we can utilize *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 trained, which differs from embedded and wrapper methods in that they do not form a feedback loop with the model (Guyon et al., 2008). Because of their independence, they tend to generalize well (Remeseiro and Bolon-Canedo, 2019).

Feature selection methods used in clinical survey data cover a broad range of techniques. 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 (Washington et al., 2019). Some of the other feature selection methods found for models involve questionnaires include wrapper models based on random forests (Niemann et al., 2020), bootstrapped feature selection (Abbas et al., 2018), principal component analysis, multicluster feature selection (Saridewi and Sari, 2020), permutation importance (Chen et al., 2023), and ReliefF (Abut et al., 2016).

One particularly useful feature selection technique 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 (Peng et al., 2005; Ramírez-Gallego et al., 2017).

Underpinning the mRMR objective function is the mutual information (MI) between classes and features, which measures the amount of shared information between two distributions. In addition to MI, Pearson's r and F-statistic p-values, can also be used in the mRMR algorithm.

2.2 Semantic Textual Similarity

Semantic textual similarity (STS) is a task where a LM is used to score the semantic similarity of two sentences, generally by evaluating the differences between embeddings generated by a model. Cosine similarity is one typical function used to measure the difference between embeddings (Reimers and Gurevych, 2019; Oniani et al., 2022).

STS scores can be produced with language models (LMs) that follow the pre-training/foundation model paradigm. These latter models are trained with a self-supervised learning task, and then modified and trained to complete a supervised task (Devlin et al., 2018; Radford et al., 2019). Pre-training is particularly useful since it results in better generalization (Erhan et al., 2010), and because it allows computationally expensive models to be reused for different tasks (Wolf et al., 2019). Some large language models (LLMs) have demonstrated capabilities at many reasoning tasks involving semantic meaning (Singhal et al., 2022; Wei et al., 2023),

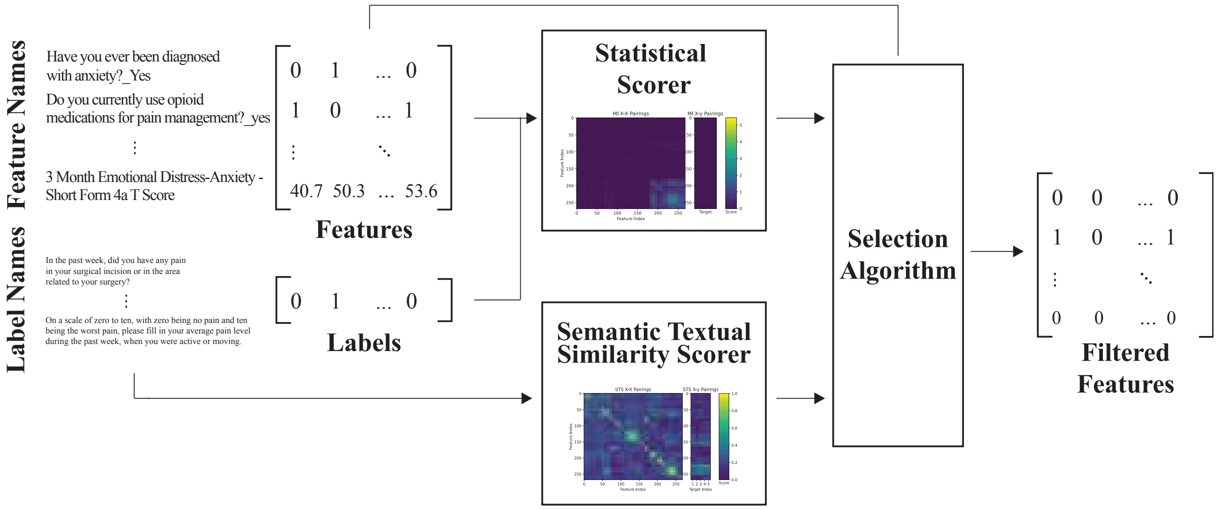

Feature Names
Have you ever been diagnosed with anxiety?_Yes
Do you currently use opioid medications for pain management?_yes
3 Month Emotional Distress-Anxiety - Short Form 4a T Score
0 1 ... 0
1 0 ... 1
40.7 50.3 ... 53.6
Features
Label Names
In the past week, did you have any pain in your surgical incision or in the area related to your surgery?
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 1 ... 0
Labels
Statistical Scorer
MI X-X Pairings
MI X-y Pairings
Feature Index
Semantic Textual Similarity Scorer
STS X-X Pairings
STS X-y Pairings
Feature Index
Target Index
Score
Selection Algorithm
0 0 ... 0
1 0 ... 1
0 0 ... 0
Filtered Features

Figure 1: Overview of the STS-based feature selection process. Feature-feature and feature-label relationships are first *scored*, using statistical algorithms and/or models fine-tuned to produce STS scores. These scores are then used to *select* the best subset of features for a downstream model.

and are highly applicable since survey features may involve complex word relationships.

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. To date, there are over 80 clinical LMs available (Wornow et al., 2023), with a diverse set of architectural and training designs. Clinical LMs can perform better than general LMs on tasks specific to the clinical domain (Alsentzer et al., 2019), and can do so more efficiently than a general-purpose models (Lehman et al., 2023).

3 Methodology

3.1 Scoring & Selection

A distinction can be made between the *scoring* of features, where we measure a feature's relationship to a target or another feature, and the *selection* of features, where we then use these scores to select the appropriate features. We evaluate the performance of STS, as well as three baseline scoring methods: MI, Pearson's r coefficients, and F-statistic p-values. We also evaluate the linear combination of them, with the coefficient for STS being a hyperparameter α, which we test over a logarithmically-spaced range of 30 values from $[10^{-2}, 10^{2}]$. For selection algorithms with these scores we evaluate selecting the top N feature-target scores, selecting feature-target scores above a given standard deviation k, and selection of N features using these scores in mRMR.

We evaluate several baseline feature selectors from the filter method category because they do not have a feedback loop from which overfitting could occur (Remeseiro and Bolon-Canedo, 2019) which makes them appropriate for small datasets. We use selection based on the best weights from a linear support vector machine (SVM) model and an XGBoost model (Chen and Guestrin, 2016) using `SelectFromModel` from `scikit-learn` (Pedregosa et al., 2011), and recursive feature elimination (RFE). The linear SVM model is tested with C over 10 logarithmically spaced values from $[10^{-2}, 1]$, while the XGBoost instance used in `SelectFromModel` has the default settings. RFE is also evaluated using a linear SVM model with $C = 1$ to select features.

For all of the aforementioned selection algorithms and the baseline algorithms we compare against, we select 40 features. For selection by standard deviations, all features with scores above $\mu + k\sigma$ are selected, where $k = 1$.

Other than formatting to support the one-hot vectorization of categorical features, the feature names are unmodified from their original text. In addition to each of the features used to assign the target label, we also use the name `persistent_pain` to the target, and average their STS scores together in evaluation. Details for how each dataset derived the target label can be found in Sections 4.1.1 and 4.1.2. The MI, and Pearson's r and F-statistic approxima-

Dataset	Type	Pairs
`bio_simlex` (Chiu et al., 2018)	Clin.	987
`bio_sim_verb` (Chiu et al., 2018)	Clin.	1,000
`mayosrs` (Pedersen et al., 2007)	Clin.	101
Total Clinical		2,088
`sts-companion` (Cer et al., 2017)	Gen.	5,289
`stsb_multi_mt;en` (May, 2021)	Gen.	5,749
Total General		11,038
Total		13,126

Table 1: Datasets used for training and fine-tuning selected models.

tion methods used in this paper are implemented from scikit-learn (Pedregosa et al., 2011).

3.2 LMs and Fine-Tuning Datasets

For generating STS scores, we evaluate the performance of several different pre-trained LMs using Hugging Face's `transformers` (Wolf et al., 2019) with the `sentence_transformers` library (Reimers and Gurevych, 2019) to fine-tune and produce STS scores. We evaluate two models trained on general vocabulary, namely `all-MiniLM-L12-v2` (Reimers and Gurevych, 2019) and `bert-base-uncased` (Devlin et al., 2018). We also evaluate the performance of models pre-trained on clinical/scientific text, including Bio_ClinicalBERT (Alsentzer et al., 2019), BioMed-RoBERTa-base (Gururangan et al., 2020), PubMedBERT (Gu et al., 2021), BioGPT (Luo et al., 2022), and GatorTron (Yang et al., 2022).

To fine-tune these models, we train the model to predict cosine similarity on one of two combined STS datasets, which we group into a general sentence-level and clinical phrase-level set. We also evaluate the performance of fine-tuning on the combination of both the general and clinical STS datasets. Table 1 highlights the composition of the selected fine-tuning datasets.

Additionally, ClinicalSTS (Xiong et al., 2020) and MedSTS (Wang et al., 2020) are two clinical sentence-pair datasets we were unable to obtain for fine-tuning. To overcome this limitation, we also evaluate the performance of a Bio_CinicalBERT model fine-tuned on just the ClinicalSTS dataset (Mulyar et al., 2019).

4 Experiments

4.1 Data

To evaluate our proposed approach, we utilize two datasets, a private dataset examining PPSP and a cohort subset from the *All of Us* gated-access dataset from the National Institutes of Health. Both datasets use the Research Electronic Data Capture (REDCap) system (Harris et al., 2009) for collecting and organizing patient survey data.

4.1.1 PPSP Dataset

The PPSP dataset is a partially complete set of participants from the *P5: Personalized Prediction of Persistent Postsurgical Pain* study (IRB #202101123) conducted at the Washington University School of Medicine in St. Louis/BJC HealthCare system. This dataset attempts to examine who will get PPSP, the phenomenon of a patient experiencing surgically-related pain for a longer duration of time than expected (Vila et al., 2020).

The dataset contains 1631 partial responses from a total 617 participants as a part of a final goal of 2,000 participants from the BJC HealthCare system. When exported from REDCap, the combination of these questionnaires results in 458 named features, ranging from demographic features to various measures of psychological and physical pain and correlated variables. Table 4 in Appendix A outlines the key characteristics of the dataset, including number of examples and general demographics.

4.1.2 *All of Us* Dataset

The *All of Us* dataset is a gated-access dataset of electronic health records, surveys, and other medical data collected by the National Institutes of Health for a broad range of biomedical research. The *All of Us* program collects extensive survey data capturing patient history.

To parallel our private dataset, we utilize the *All of Us* Registered Tier Dataset v7 and select a cohort of patients who experienced "Persistent pain following procedure" or "Pre-surgery evaluation" and use the presence of the former condition as the label for that dataset. We only include patients who have the "Contains Surveys Codes" condition, resulting in 32,631 patients.

4.2 Data & Model Preparation

To deal with missing entries in survey data, several imputation strategies are applied. For columns with numerical types of data, null entries are replaced with the mean value, and then have the L_2 norm applied to that column. Date/time types are dropped for simplicity. 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.

Classifier models tested include linear SVM, multilayer perceptron, Gaussian Naïve Bayes (Gaussian NB), and k-nearest neighbors (k-NN). For linear SVM, we test over 10 values of C logarithmically-spaced from 0.01 to 1. For k-NN, we evaluate 3, 5, and 7 neighbors, and for MLP, we evaluate tanh and ReLU activations with α logarithmically-spaced from 0.01 to 1 over 10 steps.

An 80%/20% train/test split is used for evaluating overall performance, and 5-fold flat cross-validation (CV) is employed to both select hyperparameters and evaluate the overall performance of the dataset. Nested CV is typically employed for evaluating model selection with small datasets, but experimentally may not be necessary with low numbers of hyperparameters while using specific model types, such as gradient boosted trees (Wainer and Cawley, 2021). For this reason, and due to the fact that nested CV with K outer-folds would incur a proportional increase in run-time, flat 5-fold CV is used. For randomization in NumPy and PyTorch and any of their dependencies, we use the seeds 278797835 and 424989.

To emulate the low-dimensional setting found in the PPSP dataset, the *All of Us* dataset has examples randomly dropped such that the ratio of features to examples is 1:1 (*i.e.* 1466 examples with 1466 features). The final dimensionality of the PPSP dataset after preprocessing is 617 examples with 269 features.

5 Results

We evaluated 1108 different feature selection and classifier pairings for the PPSP dataset, and 821 pairings for the *All of Us* dataset. Tables 2 and 3 show the the performance in area under the receiver-operator curve (AUROC) and area under the precision-recall curve (AUPRC) for each of the STS-based feature selectors compared to all baselines using the Gaussian NB classifier, which was the most significantly improved classifier for both datasets. Two-tailed t-test p-values between those groups, which were corrected for multiple hypotheses using the Benjamini-Yekutieli method with $\alpha = 0.05$ (Benjamini and Yekutieli, 2001; Virtanen et al., 2020; Seabold and Perktold, 2010), are also shown. Full results stratified by all classifiers tested can be found in Tables 9 to 12 in Appendix C. Detailed summaries of training results with the Gaussian NB can be found in Tables 13 and 14.

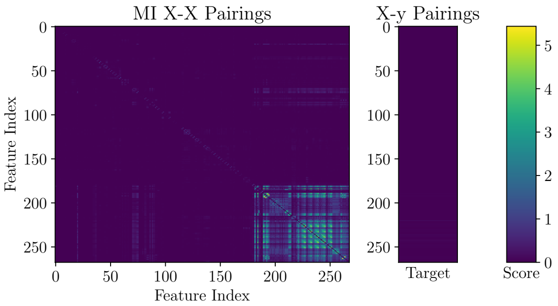


Figure 2: MI feature-feature and feature-target pairings from the PPSP dataset.

5.1 Scoring Performance

When evaluating the performance between the STS and non-STS scorers, as shown in Tables 2 and 3, we find that STS tends to slightly drop in test performance for the *All of Us* dataset and slightly improve in test performance for the PPSP dataset. We find in both cases the test-train gap improves with both STS alone or in combination with the majority of the baseline scoring methods compared to the baseline feature selectors.

We compare performance for each of these hyperparameters overall using ANOVA and Tukey's honestly significant difference (HSD) test in Tables 5 to 8 in Appendix B (Tukey, 1949; Seabold and Perktold, 2010). For training the STS scoring models, we find that the differences between a clinical and general fine-tuning dataset appear to be significant for the *All of Us* dataset but negligible with the PPSP dataset, as seen in Tables 5 and 7 in Appendix B. Similarly, for fine-tuned scoring models, we find that for the majority of pairings, there is no significant difference in model performance, as seen in Tables 6 and 8 in Appendix B.

5.2 Feature Scoring Differences

When evaluating the selected features, we see a noticeable difference between those selected using MI and STS scores. Figures 2 and 4,which shows MI, and Figures 3 and 5, which shows selected STS scores for the features and targets for the dataset, highlights why this difference exists, as the STS scores are capable of highlighting more relationships between features and targets. The performance of Pearson's r and F-statistic scores

Table 2: Results stratified by scorer measured by AUROC using Gaussian NB, with average baseline performance (μ_B), average performance for a given scorer (μ_S), average difference between baseline and scorer (Δ), and the t-test p value corrected with the Benjamini-Yekutieli method.

Hyperparameters		Test - Train AUROC				Test AUROC			
D.	Scorer	μ_B	μ_S	Δ	p	μ_B	μ_S	Δ	p
All of Us	All STS	−0.123	**−0.057**	**0.066**	**0.003**	0.628	0.578	−0.050	0.852
	F-Score & STS		−0.040	0.084	0.077		0.705	0.078	0.234
	MI & STS		**−0.059**	**0.064**	**0.009**		**0.551**	**−0.076**	**0.002**
	Pearson's r & STS		**−0.055**	**0.068**	**0.027**		0.592	−0.035	1.000
	STS		**−0.062**	**0.061**	**0.013**		**0.548**	**−0.080**	**0.002**
PPSP	All STS	−0.024	**0.060**	**0.085**	$\mathbf{7.0 \cdot 10^{-08}}$	0.742	**0.840**	**0.098**	$\mathbf{8.7 \cdot 10^{-04}}$
	F-Score & STS		**0.065**	**0.089**	$\mathbf{1.5 \cdot 10^{-05}}$		**0.839**	**0.097**	**0.024**
	MI & STS		**0.050**	**0.074**	$\mathbf{9.6 \cdot 10^{-04}}$		0.816	0.074	0.261
	Pearson's r & STS		**0.066**	**0.090**	$\mathbf{4.6 \cdot 10^{-06}}$		**0.857**	**0.115**	$\mathbf{1.7 \cdot 10^{-04}}$
	STS		**0.060**	**0.084**	$\mathbf{7.3 \cdot 10^{-06}}$		**0.849**	**0.107**	$\mathbf{1.8 \cdot 10^{-06}}$

Table 3: Results stratified by scorer measured by AUPRC using Gaussian NB.

Hyperparameters		Test - Train AUPRC				Test AUPRC			
D.	Scorer	μ_B	μ_S	Δ	p	μ_B	μ_S	Δ	p
All of Us	All STS	−0.036	**−0.011**	**0.025**	$\mathbf{4.0 \cdot 10^{-05}}$	0.033	0.027	−0.006	1.000
	F-Score & STS		−0.026	0.010	0.972		0.046	0.013	0.173
	MI & STS		**−0.007**	**0.029**	$\mathbf{7.3 \cdot 10^{-08}}$		**0.023**	**−0.010**	$\mathbf{1.9 \cdot 10^{-06}}$
	Pearson's r & STS		**−0.012**	**0.023**	**0.039**		0.030	−0.003	1.000
	STS		**−0.007**	**0.028**	$\mathbf{7.3 \cdot 10^{-08}}$		**0.023**	**−0.010**	$\mathbf{1.9 \cdot 10^{-06}}$
PPSP	All STS	0.037	**0.102**	**0.065**	**0.002**	0.296	**0.423**	**0.126**	$\mathbf{4.3 \cdot 10^{-04}}$
	F-Score & STS		**0.106**	**0.069**	**0.012**		0.410	0.114	0.056
	MI & STS		0.087	0.050	0.065		0.406	0.110	0.056
	Pearson's r & STS		**0.115**	**0.078**	$\mathbf{2.9 \cdot 10^{-05}}$		**0.440**	**0.143**	$\mathbf{1.9 \cdot 10^{-04}}$
	STS		**0.100**	**0.063**	**0.003**		**0.435**	**0.138**	$\mathbf{3.5 \cdot 10^{-05}}$

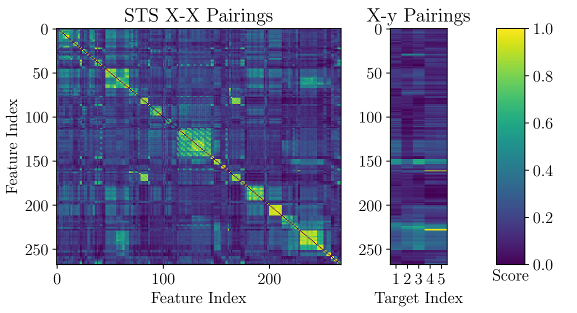


Figure 3: STS scoring with `microsoft/biogpt` fine-tuned on both the general and clinical vocab pairings from the PPSP dataset.

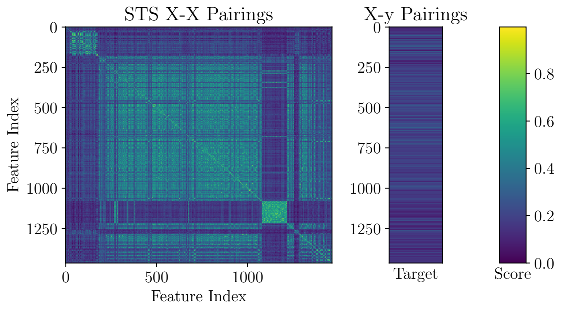


Figure 5: STS scoring with `UFNLP/gatortron-base` fine-tuned on both the general and clinical vocab pairings from the *All of Us* dataset.

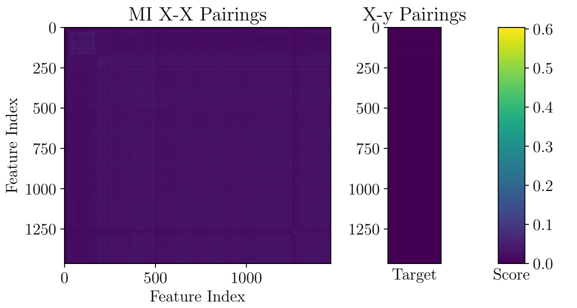


Figure 4: MI feature-feature and feature-target pairings from the *All of Us* dataset.

shows a similar trend, which can be seen in Figures 6a, 6b and 7a in Appendix D.

6 Discussion

6.1 Feature Selection Performance

Using STS to select features appears to offer significant performance benefits over traditional feature selection methods. One main reason that the usage of STS appears to work better than statistical scoring methods is that STS scores are not affected by the curse of dimensionality. As can be seen with

MI in Figures 2 and 4, complex relationships can cause individual features to share little information, and features that we would expect to be more relevant than others will be scored only marginally better than those that would not be relevant.

STS also provides a contrasting, non-correlated, source of information compared to statistical measures like MI. This can be seen from the strong contrast in highlighted regions between Figs. 2 and 3. This is particularly useful for when statistical scores that result 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. STS is clearly more able to highlight redundant regions, especially near the diagonal, and is more able to distinguish between relevant and irrelevant features than traditional statistical scores.

Inspection of the features selected with STS and statistically scored models highlights these issues. For example, in the *All of Us* dataset, when examining the top N features scored with MI, we find that the majority of features are variations of "Can't Afford Care: Skipped Med to Save Money", "Can't Afford Care: Took Less Med To Save Money", *etc.* In contrast, the top N features for PubMedBERT begin with questions such as "Are you still seeing a doctor or health care provider for a hernia?" and "Are you still seeing a doctor or health care provider for reactions to anesthesia (such as hyperthermia)?". While the first set of features is more strongly correlated to the labels in the training set, the second set of features are more semantically related to the label, and are more likely to reflect ground-truth explanations.

6.2 Implications on Survey Writing

The surveys used here were not designed in anticipation of this type of feature selection algorithm, but subsequent questionnaires—or other data with descriptive feature names—may be designed with STS-based feature selection in mind. We offer several guidelines for feature name writing.

A key consideration is that of the limitations of the LMs' vocabulary. Not all possible words or sub-words will be present, and furthering this, not all words will appear in the pre-training corpora or fine-tuning data. However, as long as either the pre-training or fine-tuning data contains references to a desired relationship, the LM should be able to produce a meaningful STS score, as the LM will have learned word relationships in its vocabulary during pre-training.

Furthering these limitations are features with hidden semantics, which may not be clearly related to the target, may be mistaken for other concepts, or may be a component of another concept. A clear example of this is the group of questions with "Color-Word Score (CW)." These particular questions have hidden semantics as they involve a phrase that is unlikely to be interpreted as the Stroop Color-Word test often used for assessing cognition (Jensen and Jr, 1966; Bjekić et al., 2018; Martinsen et al., 2014). Because this question was written for a clinician to fill out, key contextual information is absent. This particular phrase is picked up when using MI to score, but is missed when using STS for features with more direct connections to the targets. To deal with this, we suggest that phrases should be expanded to include words that connect it to the target outcome.

Another issue we do not examine in-depth is how writing style affects the outcome of the features chosen. While parts of the feature name are intrinsically linked to the outcome of the feature itself, the overall writing style (i.e., word ordering, word choice, verbosity, etc.) can be more random. Although LMs can be used to detect differences in writing style (Ríos-Toledo et al., 2022; Yamshchikov et al., 2021), the fine-tuning datasets we use here incorporate smaller positive scores for two sentence pairs sharing the same topic, and for this reason it is likely that the impact of writing style could be a smaller concern than the topics of the feature names. However, to minimize variability in selections, we still recommend adopting a uniform and detailed writing style across a set of feature names.

Although the evidence presented here suggests that there may be a negligible difference between the performance of tokenizers—as the models tested use different tokenizers—subtle differences in tokenization may change the way that semantic relationships are captured. For example, (Bostrom and Durrett, 2020) note that byte-pair encoding (Gage, 1994; Sennrich et al., 2016) has weaker performance than unigram language modeling (Kudo, 2018) with respect to morphological segmentations, implying that features tokenized with the latter algorithm may produce more meaningful STS scores.

Overall, our recommendations can be summarized as follows:

- Common and meaningful vocabulary should be prioritized to ensure an LM recognizes it.

- Acronyms should be expanded where possible, and rare or ambiguous phrases should be qualified with supporting contextual details where possible.

- Feature names should be as stylistically consistent and detailed throughout.

- Attention should be paid with the algorithm used to tokenize feature names.

6.3 Future Work

One area of future work is evaluating the performance of other measures, as the scorers evaluated here both have limitations. MI, Pearson's r, and the F-statistic are incapable of measuring the true amount of information a feature contains in context with other features, and STS only represent semantic relationships without any regard to their underlying statistical relationship.

Future work could also consider the choice of other LMs, or alternatives to fine-tuning such as adapters (Houlsby et al., 2019). Resource limitations prevented us from evaluating the performance of LMs in the billion-parameter range, although the evidence presented here suggests that there may be marginal benefits to different scoring models.

Future work regarding STS-based feature selection could also consider the use of semantic pairs that specifically rate *relevancy* and *redundancy* between pairs of questions rather than similarity. Relevant and redundant questions may not always be semantically similar, and a model fine-tuned for this may select better features than those for STS. Similarly, the performance of other selection algorithms with STS could also be evaluated.

Another potential area of future work would be to serialize the the feature selection task into a text prompt. Serialization of tabular data into a question prompt for a LLM can achieve high performance in a few-shot learning context (Hegselmann et al., 2022), and serialization of a feature selection objective may also be able to capture further semantic relationships between features.

7 Conclusion

Overall, we demonstrate that our proposed approach of using STS to score features—either alone or with statistically-based scorers—can be effective in the context of clinical survey feature selection. We discussed the intrinsic differences between STS-based and statistical scoring that result in the observed performance difference on two different datasets, as well as how different fine-tuning hyperparameters affect the performance. Finally, we suggest various considerations for writing survey questions that work with this approach, and potential areas of future research with regards to STS-based feature selection.

Limitations

One key limitation is that we only evaluate our results on one non-public and another gated-access dataset of protected health information (PHI). We were unable to find other fully-public datasets that were similar in dimensionality with descriptive feature names.

Another limitation is that many types of data were dropped from the PPSP dataset for simplicity, such as non-categorical string types. The preprocessing steps we use reduce the overall feature count from 458 features to 269 features, before any further feature selection is applied.

For the *All of Us* dataset, we did not include `microsoft/biogpt` and the ClinicalSTS fine-tuning of Bio_CinicalBERT that dataset due to package incompatibilities.

The feature selectors evaluated here are a subset of possible approaches, and future work may evaluate other baselines and STS-based selection algorithms. Further limitations that could be areas of future work are discussed further in Section 6.3.

Ethical Considerations

The data from the PPSP dataset is a part of the *P5: Personalized Prediction of Persistent Postsurgical Pain* study (IRB #202101123) performed at the Washington University School of Medicine in St. Louis/BJC HealthCare system. All code and model artifacts we release are not derived from any PHI collected in this study.

The features selected from such a model are not comprehensive, particularly in a clinical context. The example features we discuss are thus relative, and should not be used to inform any clinical decision-making.

Acknowledgments

This work was made possible through the support of the Department of Defense Congressionally Directed Medical Research Programs award W81XWH-21-1-0736, as well as the National Institute of Health through their support of the *All of Us* program. The authors would like to also thank Hanyang Liu for his valuable guidance on the final writing and publishing of this project. Finally, we would also like to thank the participants in both the *P5* study and the *All of Us* program whose partnership made this research possible.

References

Halim Abbas, Ford Garberson, Eric Glover, and Dennis P Wall. 2018. Machine learning approach for early detection of autism by combining questionnaire and home video screening. *Journal of the American Medical Informatics Association*, 25(8):1000–1007. Publisher: Oxford University Press.

Fatih Abut, Mehmet Fatih Akay, and James George. 2016. Developing new VO2max prediction models from maximal, submaximal and questionnaire variables using support vector machines combined with feature selection. *Computers in biology and medicine*, 79:182–192. Publisher: Elsevier.

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. *arXiv preprint arXiv:1904.03323*.

Yoav Benjamini and Daniel Yekutieli. 2001. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4).

Jovana Bjekić, Marko Živanović, Danka Purić, Joukje M. Oosterman, and Saša R. Filipović. 2018. Pain and executive functions: a unique relationship between Stroop task and experimentally induced pain. *Psychological Research*, 82(3):580–589.

Kaj Bostrom and Greg Durrett. 2020. Byte pair encoding is suboptimal for language model pretraining. *arXiv preprint arXiv:2004.03720*.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Yao Chen, Ke Wang, and Jian John Lu. 2023. Feature selection for driving style and skill clustering using naturalistic driving data and driving behavior questionnaire. *Accident Analysis & Prevention*, 185:107022. Publisher: Elsevier.

Billy Chiu, Sampo Pyysalo, Ivan Vulić, and Anna Korhonen. 2018. Bio-SimVerb and Bio-SimLex: wide-coverage evaluation sets of word similarity in biomedicine. *BMC bioinformatics*, 19(1):1–13. Publisher: BioMed Central.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dumitru Erhan, Aaron Courville, Yoshua Bengio, and Pascal Vincent. 2010. Why does unsupervised pre-training help deep learning? In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 201–208. JMLR Workshop and Conference Proceedings.

Philip Gage. 1994. A New Algorithm for Data Compression. *C Users J.*, 12(2):23–38. Place: USA Publisher: R & D Publications, Inc.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23. Publisher: ACM New York, NY.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of ACL*.

Isabelle Guyon, Steve Gunn, Masoud Nikravesh, and Lofti A Zadeh. 2008. *Feature extraction: foundations and applications*, volume 207. Springer.

Paul A Harris, Robert Taylor, Robert Thielke, Jonathon Payne, Nathaniel Gonzalez, and Jose G Conde. 2009. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of biomedical informatics*, 42(2):377–381. Publisher: Elsevier.

Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. 2022. TabLLM: Few-shot Classification of Tabular Data with Large Language Models. *arXiv preprint arXiv:2210.10723*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-Efficient Transfer Learning for NLP. ArXiv:1902.00751 [cs, stat].

Axthijr R Jensen and William D Rohwer Jr. 1966. The Stroop Color-Word Test: A Review. *Acta Psychologica*, 25.

Taku Kudo. 2018. Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Vasileios Lampos, Bin Zou, and Ingemar Johansson Cox. 2017. Enhancing feature selection using word embeddings: The case of flu surveillance. In *Proceedings of the 26th International Conference on World Wide Web*, pages 695–704.

Eric Lehman, Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J Smith, Zachary Ziegler, Daniel Nadler, Peter Szolovits, Alistair Johnson, and Emily Alsentzer. 2023. Do We Still Need Clinical Language Models? *arXiv preprint arXiv:2302.08091*.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6). Publisher: Oxford Academic.

Sofia Martinsen, Pär Flodin, Jonathan Berrebi, Monika Löfgren, Indre Bileviciute-Ljungar, Martin Ingvar, Peter Fransson, and Eva Kosek. 2014. Fibromyalgia Patients Had Normal Distraction Related Pain Inhibition but Cognitive Impairment Reflected in Caudate Nucleus and Hippocampus during the Stroop Color Word Test. *PLoS ONE*, 9(10):e108637.

Philip May. 2021. Machine translated multilingual STS benchmark dataset.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Andriy Mulyar, Elliot Schumacher, and Mark Dredze. 2019. semantic-text-similarity.

Uli Niemann, Petra Brueggemann, Benjamin Boecking, Birgit Mazurek, and Myra Spiliopoulou. 2020. Development and internal validation of a depression severity prediction model for tinnitus patients based on questionnaire responses and socio-demographics. *Scientific reports*, 10(1):4664. Publisher: Nature Publishing Group UK London.

David Oniani, Sonish Sivarajkumar, and Yanshan Wang. 2022. Few-Shot Learning for Clinical Natural Language Processing Using Siamese Neural Networks. *arXiv preprint arXiv:2208.14923*.

Ted Pedersen, Serguei VS Pakhomov, Siddharth Patwardhan, and Christopher G Chute. 2007. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of biomedical informatics*, 40(3):288–299. Publisher: Elsevier.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and others. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, 12:2825–2830. Publisher: JMLR. org.

Hanchuan Peng, Fuhui Long, and Chris Ding. 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 27(8):1226–1238. Publisher: IEEE.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Sergio Ramírez-Gallego, Iago Lastra, David Martínez-Rego, Verónica Bolón-Canedo, José Manuel Benítez, Francisco Herrera, and Amparo Alonso-Betanzos. 2017. Fast-mRMR: Fast minimum redundancy maximum relevance algorithm for high-dimensional big data. *International Journal of Intelligent Systems*, 32(2):134–152. Publisher: Wiley Online Library.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Beatriz Remeseiro and Veronica Bolon-Canedo. 2019. A review of feature selection methods in medical applications. *Computers in biology and medicine*, 112:103375. Publisher: Elsevier.

Germán Ríos-Toledo, Juan Pablo Francisco Posadas-Durán, Grigori Sidorov, and Noé Alejandro Castro-Sánchez. 2022. Detection of changes in literary writing style using N-grams as style markers and supervised machine learning. *PLOS ONE*, 17(7):e0267590.

Valentina Siwi Saridewi and Riri Fitri Sari. 2020. Feature selection in the human aspect of information security questionnaires using multicluster feature selection. *International Journal of Advanced Science and Technology*, 29(7 Special Issue):3484–3493. Publisher: Science and Engineering Research Support Society.

Skipper Seabold and Josef Perktold. 2010. Statsmodels: Econometric and Statistical Modeling with Python. pages 92–96, Austin, Texas.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and others. 2022. Large Language Models Encode Clinical Knowledge. *arXiv preprint arXiv:2212.13138*.

John W. Tukey. 1949. Comparing Individual Means in the Analysis of Variance. *Biometrics*, 5(2):99.

Molly R Vila, Marko S Todorovic, Cynthia Tang, Marilee Fisher, Aaron Steinberg, Beverly Field, Michael M Bottros, Michael S Avidan, and Simon Haroutounian. 2020. Cognitive flexibility and persistent post-surgical pain: the FLEXCAPP prospective observational study. *British journal of anaesthesia*, 124(5):614–622. Publisher: Elsevier.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. Van Der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul Van Mulbregt, SciPy 1.0 Contributors, Aditya Vijaykumar, Alessandro Pietro Bardelli, Alex Rothberg, Andreas Hilboll, Andreas Kloeckner, Anthony Scopatz, Antony Lee, Ariel Rokem, C. Nathan Woods, Chad Fulton, Charles Masson, Christian Häggström, Clark Fitzgerald, David A. Nicholson, David R. Hagen, Dmitrii V. Pasechnik, Emanuele Olivetti, Eric Martin, Eric Wieser, Fabrice Silva, Felix Lenders, Florian Wilhelm, G. Young, Gavin A. Price, Gert-Ludwig Ingold, Gregory E. Allen, Gregory R. Lee, Hervé Audren, Irvin Probst, Jörg P. Dietrich, Jacob Silterra, James T Webber, Janko Slavič, Joel Nothman, Johannes Buchner, Johannes Kulick, Johannes L. Schönberger, José Vinícius De Miranda Cardoso, Joscha Reimer, Joseph Harrington, Juan Luis Cano Rodríguez, Juan Nunez-Iglesias, Justin Kuczynski, Kevin Tritz, Martin Thoma, Matthew Newville, Matthias Kümmerer, Maximilian Bolingbroke, Michael Tartre, Mikhail Pak, Nathaniel J. Smith, Nikolai Nowaczyk, Nikolay Shebanov, Oleksandr Pavlyk, Per A. Brodtkorb, Perry Lee, Robert T. McGibbon, Roman Feldbauer, Sam Lewis, Sam Tygier, Scott Sievert, Sebastiano Vigna, Stefan Peterson, Surhud More, Tadeusz Pudlik, Takuya Oshima, Thomas J. Pingel, Thomas P. Robitaille, Thomas Spura, Thouis R. Jones, Tim Cera, Tim Leslie, Tiziano Zito, Tom Krauss, Utkarsh Upadhyay, Yaroslav O. Halchenko, and Yoshiki Vázquez-Baeza. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272.

Jacques Wainer and Gavin Cawley. 2021. Nested cross-validation when selecting classifiers is overzealous for most practical applications. *Expert Systems with Applications*, 182:115222. Publisher: Elsevier.

Yanshan Wang, Naveed Afzal, Sunyang Fu, Liwei Wang, Feichen Shen, Majid Rastegar-Mojarad, and Hongfang Liu. 2020. MedSTS: a resource for clinical semantic textual similarity. *Language Resources and Evaluation*, 54:57–72. Publisher: Springer.

Peter Washington, Kelley Marie Paskov, Haik Kalantarian, Nathaniel Stockham, Catalin Voss, Aaron Kline, Ritik Patnaik, Brianna Chrisman, Maya Varma, Qandeel Tariq, and others. 2019. Feature selection and dimension reduction of social autism data. In *Pacific Symposium on Biocomputing 2020*, pages 707–718.

Chengwei Wei, Yun-Cheng Wang, Bin Wang, and C-C Jay Kuo. 2023. An Overview on Language Models: Recent Developments and Outlook. *arXiv preprint arXiv:2303.05759*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and others. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Michael Wornow, Yizhe Xu, Rahul Thapa, Birju Patel, Ethan Steinberg, Scott Fleming, Michael A Pfeffer, Jason Fries, and Nigam H Shah. 2023. The shaky foundations of large language models and foundation models for electronic health records. *npj Digital Medicine*, 6(1):135. Publisher: Nature Publishing Group UK London.

Ying Xiong, Shuai Chen, Qingcai Chen, Jun Yan, Buzhou Tang, and others. 2020. Using character-level and entity-level representations to enhance bidirectional encoder representation from transformers-based clinical semantic textual similarity model: ClinicalSTS modeling study. *JMIR Medical Informatics*, 8(12):e23357. Publisher: JMIR Publications Inc., Toronto, Canada.

Ivan P. Yamshchikov, Viacheslav Shibaev, Nikolay Khlebnikov, and Alexey Tikhonov. 2021. Style-transfer and Paraphrase: Looking for a Sensible Semantic Similarity Metric. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14213–14220.

Xi Yang, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Mona G Flores, Ying Zhang, Tanja Magoc, and others. 2022. GatorTron: A Large Language Model for Clinical Natural Language Processing. *medRxiv*, pages 2022–02. Publisher: Cold Spring Harbor Laboratory Press.

A Private PPSP Dataset Demographics

The demographics of the PPSP dataset are reported in Table 4.

Name	Value
PPSP Characteristics	
Completed Cases	617
Responses by Individual (mean)	2.64
Responses by Individual (std. dev.)	1.54
Responses by Individual (max)	10
Positive Cases	25
Negative Cases	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
(No Answer)	7
Age	
Age (min)	19
Age (mean)	52.47
Age (std. dev.)	13.52
Age (max)	75

Table 4: Demographics of the partial PPSP dataset.

B Performance By Feature Selector Hyperparameter

Tables 5 and 6 highlights the performance of all classifiers tested as fine-tuning dataset, feature selector, and STS scoring models are changed, respectively, for the *All of Us* dataset.

Tables 7 and 8 highlights the performance of all classifiers tested as fine-tuning dataset, feature selector, and STS scoring models are changed, respectively, for the PPSP dataset.

C Performance By Classifier

Tables 13 and 14 summarizes the performance of individual configurations of Gaussian NB for the *All of Us* and PPSP dataset respectively.

D Baseline Feature Scoring Performance

Figures 6a, 6b, 7a and 7b show the feature and target scorings using Pearson's r and F-statistic for both the *All of Us* and Pearson's r datasets.

Table 5: Results for fine-tuning dataset grouped by test AUROC for the *All of Us* dataset. One-way ANOVA $p = 0.011$, significant Tukey HSD results highlighted.

Fine-Tuning Dataset	Mean	Std. Dev.	(1)	(2)	(3)
Clinical Pairs (1)	0.572	0.091	-	**0.026**	**0.024**
Combined Pairs (2)	0.594	0.095	**0.026**	-	1.000
General Pairs (3)	0.594	0.101	**0.024**	1.000	-

Table 6: Results for scoring model grouped by test AUROC for the *All of Us* dataset. One-way ANOVA $p = 0.001$, significant Tukey HSD results highlighted.

Scoring Model	Mean	Std. Dev.	(1)	(2)	(3)	(4)	(5)	(6)
Bio_ClinicalBERT (1)	0.569	0.080	-	**0.012**	0.966	1.000	0.369	**0.033**
bert-base-uncased (2)	0.608	0.093	**0.012**	-	0.121	**0.010**	0.747	1.000
all-MiniLM-L12-v2 (3)	0.578	0.097	0.966	0.121	-	0.953	0.862	0.243
biomed_roberta_base (4)	0.568	0.099	1.000	**0.010**	0.953	-	0.332	**0.027**
gatortron-base (5)	0.592	0.101	0.369	0.747	0.862	0.332	-	0.903
PubMedBERT (6)	0.604	0.100	**0.033**	1.000	0.243	**0.027**	0.903	-

Table 7: Results for fine-tuning dataset grouped by test AUROC for the PPSP dataset. One-way ANOVA $p = 0.027$, significant Tukey HSD results highlighted.

Fine-Tuning Dataset	Mean	Std. Dev.	(1)	(2)	(3)	(4)
Combined Pairs (1)	0.709	0.107	-	0.994	0.061	0.457
Clinical Pairs (2)	0.707	0.105	0.994	-	**0.043**	0.308
ClinicalSTS (3)	0.749	0.089	0.061	**0.043**	-	0.292
General Pairs (4)	0.720	0.101	0.457	0.308	0.292	-

Table 8: Results for scoring model grouped by test AUROC for the PPSP dataset. One-way ANOVA $p = 0.799$, significant Tukey HSD results highlighted.

Scoring Model	Mean	Std. Dev.	(1)	(2)	(3)	(4)	(5)	(6)	(7)
PubMedBERT (1)	0.710	0.106	-	0.998	1.000	0.996	1.000	0.941	0.999
all-MiniLM-L12-v2 (2)	0.717	0.097	0.998	-	1.000	1.000	1.000	0.999	0.943
bert-base-uncased (3)	0.712	0.098	1.000	1.000	-	0.999	1.000	0.973	0.996
biomed_roberta_base (4)	0.718	0.109	0.996	1.000	0.999	-	1.000	1.000	0.923
gatortron-base (5)	0.713	0.107	1.000	1.000	1.000	1.000	-	0.987	0.990
biogpt (6)	0.723	0.100	0.941	0.999	0.973	1.000	0.987	-	0.701
Bio_ClinicalBERT (7)	0.705	0.111	0.999	0.943	0.996	0.923	0.990	0.701	-

Table 9: t-test results for scorer grouped by test - train AUROC.

Selection Hyperparameters		Test - Train AUROC			
Classifier	Scorer	μ_B	μ_S	Δ	p
		All of Us			
All	All STS	-0.176	$\mathbf{-0.085}$	$\mathbf{0.091}$	$\mathbf{5.9 \cdot 10^{-09}}$
	F-Score & STS		$\mathbf{-0.057}$	$\mathbf{0.120}$	$\mathbf{5.9 \cdot 10^{-09}}$
	MI & STS		$\mathbf{-0.087}$	$\mathbf{0.089}$	$\mathbf{8.0 \cdot 10^{-07}}$
	Pearson's r & STS		$\mathbf{-0.105}$	$\mathbf{0.072}$	$\mathbf{2.6 \cdot 10^{-04}}$
	STS		$\mathbf{-0.081}$	$\mathbf{0.095}$	$\mathbf{1.5 \cdot 10^{-07}}$
Linear SVM	All STS	-0.137	$\mathbf{-0.067}$	$\mathbf{0.071}$	$\mathbf{0.025}$
	F-Score & STS		$\mathbf{-0.011}$	$\mathbf{0.127}$	$\mathbf{3.0 \cdot 10^{-04}}$
	MI & STS		-0.074	0.064	0.183
	Pearson's r & STS		-0.098	0.039	0.645
	STS		-0.066	0.072	0.085
MLP	All STS	-0.226	$\mathbf{-0.120}$	$\mathbf{0.106}$	$\mathbf{0.009}$
	F-Score & STS		$\mathbf{-0.087}$	$\mathbf{0.139}$	$\mathbf{0.004}$
	MI & STS		$\mathbf{-0.119}$	$\mathbf{0.107}$	$\mathbf{0.025}$
	Pearson's r & STS		-0.150	0.076	0.196
	STS		$\mathbf{-0.114}$	$\mathbf{0.112}$	$\mathbf{0.019}$
Gaussian NB	All STS	-0.123	$\mathbf{-0.057}$	$\mathbf{0.066}$	$\mathbf{0.003}$
	F-Score & STS		-0.040	0.084	0.077
	MI & STS		$\mathbf{-0.059}$	$\mathbf{0.064}$	$\mathbf{0.009}$
	Pearson's r & STS		$\mathbf{-0.055}$	$\mathbf{0.068}$	$\mathbf{0.027}$
	STS		$\mathbf{-0.062}$	$\mathbf{0.061}$	$\mathbf{0.013}$
k-NN	All STS	-0.214	$\mathbf{-0.095}$	$\mathbf{0.119}$	$\mathbf{4.2 \cdot 10^{-04}}$
	F-Score & STS		$\mathbf{-0.081}$	$\mathbf{0.133}$	$\mathbf{0.003}$
	MI & STS		$\mathbf{-0.095}$	$\mathbf{0.119}$	$\mathbf{0.003}$
	Pearson's r & STS		$\mathbf{-0.115}$	$\mathbf{0.099}$	$\mathbf{0.009}$
	STS		$\mathbf{-0.083}$	$\mathbf{0.131}$	$\mathbf{0.002}$
		PPSP			
All	All STS	-0.108	-0.070	0.038	0.569
	F-Score & STS		-0.091	0.016	1.000
	MI & STS		-0.085	0.022	1.000
	Pearson's r & STS		-0.056	0.051	0.171
	STS		-0.046	0.062	0.066
Linear SVM	All STS	-0.031	0.008	0.039	0.187
	F-Score & STS		-0.010	0.021	1.000
	MI & STS		-0.012	0.019	1.000
	Pearson's r & STS		0.022	0.053	0.108
	STS		$\mathbf{0.032}$	$\mathbf{0.063}$	$\mathbf{0.017}$
MLP	All STS	-0.091	-0.074	0.017	1.000
	F-Score & STS		-0.121	-0.030	1.000
	MI & STS		-0.113	-0.021	1.000
	Pearson's r & STS		-0.056	0.036	1.000
	STS		-0.005	0.086	0.053
Gaussian NB	All STS	-0.024	$\mathbf{0.060}$	$\mathbf{0.085}$	$\mathbf{7.0 \cdot 10^{-08}}$
	F-Score & STS		$\mathbf{0.065}$	$\mathbf{0.089}$	$\mathbf{1.5 \cdot 10^{-05}}$
	MI & STS		$\mathbf{0.050}$	$\mathbf{0.074}$	$\mathbf{9.6 \cdot 10^{-04}}$
	Pearson's r & STS		$\mathbf{0.066}$	$\mathbf{0.090}$	$\mathbf{4.6 \cdot 10^{-06}}$
	STS		$\mathbf{0.060}$	$\mathbf{0.084}$	$\mathbf{7.3 \cdot 10^{-06}}$
k-NN	All STS	-0.284	-0.273	0.011	1.000
	F-Score & STS		-0.299	-0.015	1.000
	MI & STS		-0.267	0.017	1.000
	Pearson's r & STS		-0.257	0.027	1.000
	STS		-0.269	0.014	1.000

Table 10: t-test results for feature selector grouped by test AUROC.

Selection Hyperparameters		Test AUROC			
Classifier	Scorer	μ_B	μ_S	Δ	p
		All of Us			
All	All STS	0.595	0.586	-0.009	1.000
	F-Score & STS		0.589	-0.006	1.000
	MI & STS		0.581	-0.014	1.000
	Pearson's r & STS		0.598	0.003	1.000
	STS		0.579	-0.016	1.000
Linear SVM	All STS	0.653	0.648	-0.005	1.000
	F-Score & STS		0.627	-0.026	1.000
	MI & STS		0.648	-0.005	1.000
	Pearson's r & STS		0.663	0.010	1.000
	STS		0.647	-0.006	1.000
MLP	All STS	0.609	0.615	0.006	1.000
	F-Score & STS		0.576	-0.033	1.000
	MI & STS		0.622	0.013	1.000
	Pearson's r & STS		0.631	0.022	1.000
	STS		0.618	0.009	1.000
Gaussian NB	All STS	0.628	0.578	-0.050	0.852
	F-Score & STS		0.705	0.078	0.234
	MI & STS		**0.551**	$\mathbf{-0.076}$	**0.002**
	Pearson's r & STS		0.592	-0.035	1.000
	STS		**0.548**	$\mathbf{-0.080}$	**0.002**
k-NN	All STS	0.494	0.504	0.010	1.000
	F-Score & STS		0.506	0.012	1.000
	MI & STS		0.504	0.010	1.000
	Pearson's r & STS		0.506	0.012	1.000
	STS		0.503	0.009	1.000
		PPSP			
All	All STS	0.688	0.714	0.026	0.773
	F-Score & STS		0.703	0.015	1.000
	MI & STS		0.700	0.012	1.000
	Pearson's r & STS		0.730	0.042	0.160
	STS		0.722	0.035	0.261
Linear SVM	All STS	0.729	0.710	-0.019	1.000
	F-Score & STS		0.718	-0.011	1.000
	MI & STS		0.689	-0.040	0.261
	Pearson's r & STS		0.726	-0.003	1.000
	STS		0.706	-0.023	1.000
MLP	All STS	0.659	0.658	-0.001	1.000
	F-Score & STS		0.638	-0.021	1.000
	MI & STS		0.645	-0.014	1.000
	Pearson's r & STS		0.668	0.009	1.000
	STS		0.680	0.020	1.000
Gaussian NB	All STS	0.742	**0.840**	**0.098**	$\mathbf{8.7 \cdot 10^{-04}}$
	F-Score & STS		**0.839**	**0.097**	**0.024**
	MI & STS		0.816	0.074	0.261
	Pearson's r & STS		**0.857**	**0.115**	$\mathbf{1.7 \cdot 10^{-04}}$
	STS		**0.849**	**0.107**	$\mathbf{1.8 \cdot 10^{-06}}$
k-NN	All STS	0.620	0.647	0.027	1.000
	F-Score & STS		0.616	-0.004	1.000
	MI & STS		0.649	0.029	1.000
	Pearson's r & STS		0.667	0.047	0.463
	STS		0.655	0.035	1.000

Table 11: t-test results for scorer grouped by test - train AUPRC.

Selection Hyperparameters		Test - Train AUPRC			
Classifier	Scorer	μ_B	μ_S	Δ	p
		All of Us			
All	All STS	-0.183	$\mathbf{-0.076}$	$\mathbf{0.107}$	$\mathbf{1.2 \cdot 10^{-08}}$
	F-Score & STS		$\mathbf{-0.083}$	$\mathbf{0.100}$	$\mathbf{1.3 \cdot 10^{-04}}$
	MI & STS		$\mathbf{-0.061}$	$\mathbf{0.122}$	$\mathbf{1.6 \cdot 10^{-08}}$
	Pearson's r & STS		$\mathbf{-0.105}$	$\mathbf{0.078}$	$\mathbf{0.002}$
	STS		$\mathbf{-0.058}$	$\mathbf{0.125}$	$\mathbf{1.2 \cdot 10^{-08}}$
Linear SVM	All STS	-0.228	$\mathbf{-0.106}$	$\mathbf{0.123}$	$\mathbf{0.002}$
	F-Score & STS		$\mathbf{-0.094}$	$\mathbf{0.134}$	$\mathbf{0.009}$
	MI & STS		$\mathbf{-0.084}$	$\mathbf{0.145}$	$\mathbf{0.002}$
	Pearson's r & STS		-0.160	0.069	0.328
	STS		$\mathbf{-0.082}$	$\mathbf{0.147}$	$\mathbf{0.002}$
MLP	All STS	-0.345	$\mathbf{-0.141}$	$\mathbf{0.205}$	$\mathbf{3.5 \cdot 10^{-05}}$
	F-Score & STS		$\mathbf{-0.136}$	$\mathbf{0.209}$	$\mathbf{0.002}$
	MI & STS		$\mathbf{-0.121}$	$\mathbf{0.224}$	$\mathbf{1.3 \cdot 10^{-04}}$
	Pearson's r & STS		$\mathbf{-0.190}$	$\mathbf{0.155}$	$\mathbf{0.010}$
	STS		$\mathbf{-0.113}$	$\mathbf{0.232}$	$\mathbf{9.2 \cdot 10^{-05}}$
Gaussian NB	All STS	-0.036	$\mathbf{-0.011}$	$\mathbf{0.025}$	$\mathbf{4.0 \cdot 10^{-05}}$
	F-Score & STS		-0.026	0.010	0.972
	MI & STS		$\mathbf{-0.007}$	$\mathbf{0.029}$	$\mathbf{7.3 \cdot 10^{-08}}$
	Pearson's r & STS		$\mathbf{-0.012}$	$\mathbf{0.023}$	$\mathbf{0.039}$
	STS		$\mathbf{-0.007}$	$\mathbf{0.028}$	$\mathbf{7.3 \cdot 10^{-08}}$
k-NN	All STS	-0.110	$\mathbf{-0.041}$	$\mathbf{0.069}$	$\mathbf{3.5 \cdot 10^{-05}}$
	F-Score & STS		$\mathbf{-0.047}$	$\mathbf{0.063}$	$\mathbf{0.037}$
	MI & STS		$\mathbf{-0.030}$	$\mathbf{0.080}$	$\mathbf{8.1 \cdot 10^{-06}}$
	Pearson's r & STS		$\mathbf{-0.059}$	$\mathbf{0.051}$	$\mathbf{0.039}$
	STS		$\mathbf{-0.029}$	$\mathbf{0.081}$	$\mathbf{9.1 \cdot 10^{-06}}$
		PPSP			
All	All STS	-0.097	-0.051	0.045	0.250
	F-Score & STS		-0.075	0.022	1.000
	MI & STS		-0.064	0.032	0.941
	Pearson's r & STS		-0.037	0.060	0.096
	STS		$\mathbf{-0.030}$	$\mathbf{0.067}$	$\mathbf{0.030}$
Linear SVM	All STS	-0.041	$\mathbf{0.011}$	$\mathbf{0.052}$	$\mathbf{0.023}$
	F-Score & STS		-0.016	0.024	1.000
	MI & STS		0.006	0.047	0.297
	Pearson's r & STS		$\mathbf{0.027}$	$\mathbf{0.068}$	$\mathbf{0.035}$
	STS		$\mathbf{0.029}$	$\mathbf{0.070}$	$\mathbf{0.021}$
MLP	All STS	-0.154	-0.087	0.067	0.458
	F-Score & STS		-0.132	0.022	1.000
	MI & STS		-0.135	0.019	1.000
	Pearson's r & STS		-0.059	0.095	0.264
	STS		$\mathbf{-0.023}$	$\mathbf{0.131}$	$\mathbf{0.021}$
Gaussian NB	All STS	0.037	$\mathbf{0.102}$	$\mathbf{0.065}$	$\mathbf{0.002}$
	F-Score & STS		$\mathbf{0.106}$	$\mathbf{0.069}$	$\mathbf{0.012}$
	MI & STS		0.087	0.050	0.065
	Pearson's r & STS		$\mathbf{0.115}$	$\mathbf{0.078}$	$\mathbf{2.9 \cdot 10^{-05}}$
	STS		$\mathbf{0.100}$	$\mathbf{0.063}$	$\mathbf{0.003}$
	All STS		-0.232	-0.003	1.000
k-NN	F-Score & STS	-0.229	-0.256	-0.026	0.627
	MI & STS		-0.215	0.014	1.000
	Pearson's r & STS		-0.232	-0.003	1.000
	STS		-0.224	0.005	1.000

Table 12: t-test results for scorer grouped by test AUPRC.

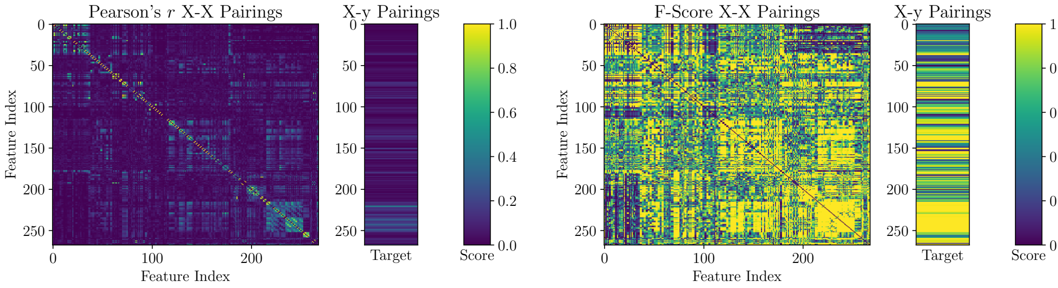
Selection Hyperparameters		Test AUPRC			
Classifier	Scorer	μ_B	μ_S	Δ	p
All of Us					
All	All STS	0.036	0.034	-0.001	1.000
	F-Score & STS		0.034	-0.002	1.000
	MI & STS		0.033	-0.002	1.000
	Pearson's r & STS		0.037	0.002	1.000
	STS		0.033	-0.002	1.000
Linear SVM	All STS	0.049	0.045	-0.004	1.000
	F-Score & STS		0.038	-0.011	1.000
	MI & STS		0.046	-0.003	1.000
	Pearson's r & STS		0.050	0.001	1.000
	STS		0.044	-0.005	1.000
MLP	All STS	0.039	0.040	0.001	1.000
	F-Score & STS		0.035	-0.004	1.000
	MI & STS		0.040	0.001	1.000
	Pearson's r & STS		0.043	0.004	1.000
	STS		0.042	0.003	1.000
Gaussian NB	All STS	0.033	0.027	-0.006	1.000
	F-Score & STS		0.046	0.013	0.173
	MI & STS		$\mathbf{0.023}$	$\mathbf{-0.010}$	$\mathbf{1.9 \cdot 10^{-06}}$
	Pearson's r & STS		0.030	-0.003	1.000
	STS		$\mathbf{0.023}$	$\mathbf{-0.010}$	$\mathbf{1.9 \cdot 10^{-06}}$
k-NN	All STS	0.021	0.025	0.003	1.000
	F-Score & STS		0.022	0.001	1.000
	MI & STS		0.024	0.003	1.000
	Pearson's r & STS		0.027	0.005	1.000
	STS		0.025	0.004	1.000
PPSP					
All	All STS	0.246	0.273	0.027	0.673
	F-Score & STS		0.253	0.007	1.000
	MI & STS		0.267	0.021	1.000
	Pearson's r & STS		0.287	0.041	0.190
	STS		0.285	0.039	0.190
Linear SVM	All STS	0.271	0.252	-0.019	0.838
	F-Score & STS		0.245	-0.026	0.695
	MI & STS		0.244	-0.027	0.673
	Pearson's r & STS		0.271	-0.000	1.000
	STS		0.247	-0.024	0.695
MLP	All STS	0.236	0.214	-0.021	0.915
	F-Score & STS		0.191	-0.045	0.134
	MI & STS		0.209	-0.027	0.673
	Pearson's r & STS		0.228	-0.008	1.000
	STS		0.230	-0.006	1.000
Gaussian NB	All STS	0.296	$\mathbf{0.423}$	$\mathbf{0.126}$	$\mathbf{4.3 \cdot 10^{-04}}$
	F-Score & STS		0.410	0.114	0.056
	MI & STS		0.406	0.110	0.056
	Pearson's r & STS		$\mathbf{0.440}$	$\mathbf{0.143}$	$\mathbf{1.9 \cdot 10^{-04}}$
	STS		$\mathbf{0.435}$	$\mathbf{0.138}$	$\mathbf{3.5 \cdot 10^{-05}}$
k-NN	All STS	0.182	0.204	0.023	1.000
	F-Score & STS		0.167	-0.015	1.000
	MI & STS		0.211	0.029	0.954
	Pearson's r & STS		0.211	0.030	0.673
	STS		0.229	0.047	0.673

Table 13: Results for Gaussian NB with baseline feature selection methods and using UFNLP/gatortron-base to score features for the *All of Us* dataset. $*$ is best test AUROC for given feature selector among ours. † is smallest test-train AUROC difference for given feature selector among ours.

Selection Hyperparameters			AUROC			AUPRC		
Feature Selector	Scorer	FT Dataset	Test	Train	Δ	Test	Train	Δ
mRMR	Pearson's r & STS	General Pairs	**0.823**	0.761	0.062	**0.076**	0.093	-0.017
Top N†	Pearson's r & STS	General Pairs	0.580	0.591	-0.011	0.024	0.028	-0.004
Top N	Pearson's r & STS	Combined Pairs	0.628	0.718	-0.090	0.031	0.071	-0.040
mRMR†	Pearson's r & STS	Combined Pairs	0.540	0.556	-0.016	0.022	0.026	-0.004
Std. Dev.	Pearson's r & STS	Clinical Pairs	0.613	0.711	-0.099	0.027	0.039	-0.013
Std. Dev.†	MI & STS	Clinical Pairs	0.568	0.664	-0.096	0.024	0.034	-0.010
Std. Dev.†	STS	Clinical Pairs	0.568	0.664	-0.096	0.024	0.034	-0.010
mRMR	Pearson's r	-	0.745	0.743	**0.003**	0.045	0.068	-0.023
Top N	MI	-	0.689	0.647	0.042	0.034	0.036	**-0.002**
Std. Dev.	MI	-	0.685	0.688	-0.003	0.033	0.038	-0.004
mRMR	MI	-	0.670	0.748	-0.079	0.042	0.067	-0.026
SFM-XGBoost	-	-	0.654	0.800	-0.147	0.056	**0.165**	-0.108
Top N	Pearson's r	-	0.648	0.759	-0.111	0.029	0.086	-0.056
Identity	-	-	0.602	0.790	-0.188	0.026	0.053	-0.027
RFE	-	-	0.574	**0.830**	-0.256	0.026	0.085	-0.059
mRMR	F-statistic	-	0.552	0.758	-0.206	0.027	0.067	-0.041
SFM-LinearSVM	-	-	0.547	0.775	-0.228	0.023	0.052	-0.029
Std. Dev.	Pearson's r	-	0.536	0.720	-0.183	0.022	0.040	-0.018

Table 14: Results for Gaussian NB with baseline feature selection methods and using emilyalsentzer/Bio_ClinicalBERT to score features for the PPSP dataset.

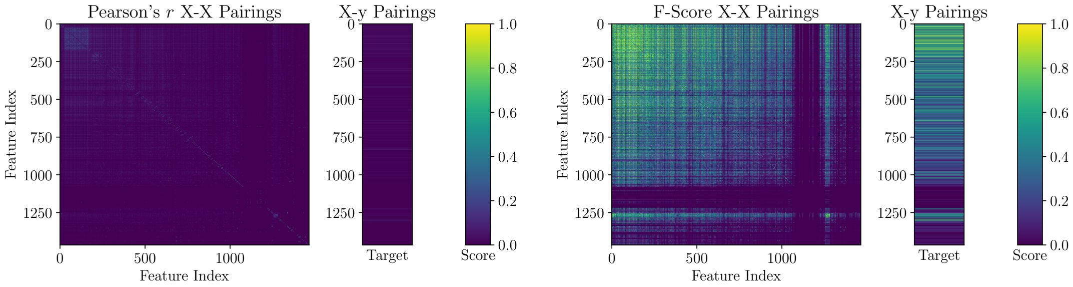
Selection Hyperparameters			AUROC			AUPRC		
Feature Selector	Scorer	FT Dataset	Test	Train	Δ	Test	Train	Δ
Std. Dev.	F-statistic & STS	General Pairs	**0.914**	0.828	0.085	0.499	0.309	**0.190**
Std. Dev.†	STS	General Pairs	0.840	0.811	0.029	0.458	0.329	0.129
mRMR†	F-statistic & STS	General Pairs	0.615	0.616	**-0.001**	0.140	0.117	0.022
Top N	Pearson's r & STS	Combined Pairs	0.906	0.798	0.109	0.495	0.373	0.122
Top N†	MI & STS	Clinical Pairs	0.852	0.776	0.076	0.443	0.339	0.104
mRMR	Pearson's r & STS	Clinical Pairs	0.841	0.793	0.048	0.333	0.269	0.064
Top N	F-statistic	-	0.907	0.814	0.093	0.482	0.324	0.157
Top N	Pearson's r	-	0.907	0.814	0.093	0.482	0.324	0.157
Std. Dev.	Pearson's r	-	0.845	0.764	0.080	0.291	0.196	0.095
SFM-XGBoost	-	-	0.817	0.828	-0.011	0.434	0.391	0.043
mRMR	F-statistic	-	0.790	0.852	-0.062	0.393	0.415	-0.022
mRMR	Pearson's r	-	0.770	0.753	0.017	0.220	0.178	0.043
Top N	MI	-	0.738	0.879	-0.140	0.384	0.396	-0.012
Std. Dev.	MI	-	0.737	**0.879**	-0.142	0.383	0.391	-0.008
mRMR	MI	-	0.719	0.736	-0.017	0.219	0.190	0.029
RFE	-	-	0.622	0.623	-0.002	0.142	0.119	0.022
Std. Dev.	F-statistic	-	0.615	0.616	-0.002	0.140	0.117	0.022
SFM-LinearSVM	-	-	0.598	0.758	-0.160	0.153	0.203	-0.050
Identity	-	-	0.583	0.645	-0.062	0.130	0.126	**0.005**



(a) Pearson's r from the PPSP dataset.

(b) F-statistic from the PPSP dataset.

Figure 6: Feature-feature and feature-target pairings from the PPSP.



(a) Pearson's r from the *All of Us* dataset.

(b) F-statistic from the *All of Us* dataset.

Figure 7: Feature-feature and feature-target pairings from the *All of Us* dataset.