

Semantic Feature Clustering Using LLM Embeddings for Label-Dependent Feature Selection

Shikha Choudhary*, Bhawna Saxena

Department of Computer Science and Engineering and IT, Jaypee Institute of Information Technology, Noida, India

Abstract—Most feature selection methods assess feature relevance using individual scoring criteria followed by a separate selection step. Although clustering-based approaches may at times capture relationships among features, relevance and redundancy are not always addressed in unison. Consequently, redundant features may receive high relevance scores and get selected, possibly leading to less informative feature subsets. This study proposes Large Language Model - embedding-based Semantic Feature Clustering (LLM-SFC), a two-level feature selection pipeline that combines semantic clustering with label-dependent feature selection. In the first level, each feature, expressed as a brief label-agnostic natural-language description, is transformed into a numerical representation using a pretrained sentence-embedding language model. Subsequently, semantic clustering of these embeddings is performed. In the second level, a label-dependent statistical criterion selects one representative feature from each cluster. This reduces redundancy while retaining features relevant to the target label. The proposed pipeline evaluates two clustering methods, Agglomerative Clustering and DBSCAN, and three label-dependent selection statistics. Experiments on three datasets from diverse domains show that agglomerative clustering reduced the feature set by 22.8–55.9% while improving mean accuracy. Statistical significance is observed for 40 of 63 classifier-statistic combinations with no significant losses anywhere. DBSCAN performed comparably when reduction was moderate, but performance dropped under aggressive reduction. Overall, the results demonstrate that semantic clustering can effectively reduce redundancy while preserving or improving predictive performance.

Keywords—Feature selection; semantic clustering; large language model embeddings; label-dependent; dimensionality reduction

I. INTRODUCTION

Modern datasets often expand in width before they increase in depth, as sensors introduce new channels, forms add additional fields, and logs keep adding more columns over time. Once a model reaches a certain level, it starts using more resources to handle unnecessary information rather than focusing on learning from meaningful data. This makes feature selection an increasingly important step in creating efficient machine learning (ML) pipelines, as choosing the most suitable one for a particular dataset is now a complex task. A survey of two decades of work pertaining to feature selection establishes that due to the large number of methods available, choosing the most suitable method has itself become a challenge [1]. Selected features can at times vary unstably across runs, even when accuracy looks fine [2]. Established feature selection methods span filter, wrapper, and hybrid approaches, such as the ANOVA F-statistic (ANOVA-F), Mutual Information (MI),

Random Forest Importance (RFI) [3], and Minimum Redundancy Maximum Relevance (mRMR) [4]. Most of them, however, treat the feature set as an unorganized list, with no built-in way to recognize when two features convey the same underlying information.

Review of existing literature reveals that existing approaches often assess feature relevance and redundancy as separate steps. Many approaches first assign relevance scores to individual features and then select a subset. Unsupervised approaches may at times consider underlying relationships without explicitly addressing feature redundancy. As a consequence, highly correlated or redundant features might get considered during selection, possibly leading to less informative feature subsets. This study addresses the identified gap by examining the differences that occur when features are clustered based on their meaning prior to applying any scoring method. This work proposes Large Language Model-embedding-based Semantic Feature Clustering (LLM-SFC), a two-level pipeline for label-dependent feature selection. In the first level, which is label-agnostic, each feature is expressed as a short natural-language description of what it represents, and this description is converted into a numerical representation using a pretrained transformer-based language embedding model. Thereafter, features are grouped by semantic similarity using a clustering algorithm, i.e., features that end up near each other in this embedded space are grouped. In the label-dependent second level, only one representative feature is selected from each cluster generated in the previous level, using a label-dependent selection statistic. The final feature set obtained has reduced dimensionality as well as less redundancy. The proposed LLM-SFC pipeline employs two varied clustering algorithms for grouping by semantic similarity: Agglomerative and Density-Based Spatial Clustering of Applications with Noise (DBSCAN). Further, in the second level, three prevalent label-dependent selection statistics, namely RFI, ANOVA-F, and MI, are used for selecting the representative feature from each cluster.

This work makes the following research contributions:

- Proposes LLM-embedding-based Semantic Feature Clustering (LLM-SFC), a two-level feature selection pipeline that first clusters features by the semantic similarity of their descriptions, before any data value or label is used, and then applies a label-dependent statistic to select features from within each cluster rather than across the full feature set. Unlike existing feature-clustering methods, which group features by statistics computed from the data, LLM-SFC groups them by meaning.

*Corresponding author.

- To establish a more reliable default, evaluation of the proposed LLM-SFC pipeline is done using two varied clustering methods, Agglomerative (hierarchical-based) and DBSCAN (density-based), and three different label-dependent selection statistics, RFI, ANOVA-F, and MI.
- Assesses the generalization capability of the proposed LLM-SFC pipeline across heterogeneous data domains.
- To the best of our knowledge, this study presents a novel feature selection pipeline wherein semantic clustering of features is first done based on language understanding, and then one representative feature is selected from each cluster using a label-dependent selection statistic.

II. RELATED WORK

A. Relevance and Redundancy in Feature Selection

Feature selection methods are commonly grouped into filter, wrapper, and embedded families [1]. Filter methods such as ANOVA-F and MI score each feature against the target independently. Wrapper and embedded methods such as RFI [3] derive importance from a trained model. Selected subsets can vary across runs even when accuracy is stable [2], and impurity-based importance degrades when inputs are correlated, because importance is shared among redundant features [21]. Redundancy-aware filters address this directly. mRMR [27] adds features greedily, rewarding relevance to the target while penalizing mutual information with features already selected, and later variants refine this trade-off over feature subsets [4]. However, in all of these methods, redundancy is estimated from data through pairwise statistics computed across the full feature set.

B. Clustering-Based Feature Selection

Clustering has also been widely used for feature selection, and existing work follows two broad directions. The first clusters the features themselves. FAST builds a minimum spanning tree over features weighted by symmetric uncertainty, partitions it into clusters, and retains from each cluster the feature most strongly related to the target [28]. The second direction, largely unsupervised, clusters samples and uses the resulting structure to guide selection. Multi-view selection is combined with graph learning [8], non-negative matrix factorization is guided by soft cluster labels [9], adaptive neighbors are weighted and projected [10], and local clustering is coupled with selection under a shared constraint [11]. Clustering itself has been applied across many domains [5–7]. Nevertheless, in both directions, groups are formed from numeric data values. Which features end up together therefore depends on the sample at hand, and the grouping must be recomputed whenever the data change.

C. Language-Model and Embedding-Based Approaches

More recent work treats features as text. LLM-Select shows that a language model prompted with feature names can rank features competitively with data-driven methods without accessing training data [12]. TabLLM shows that tabular rows serialized as natural language support few-shot classification [13]. An existing study has investigated the utilization of LLM embeddings as inputs to deep tabular models [15]. LLMs have also been surveyed across machine learning workflow

construction [14]. These methods score or represent features individually. Yet, none group features by meaning before selection, so redundancy among semantically related features is left unaddressed.

To address the gaps identified from the existing literature, this study proposes LLM-SFC, which forms clusters based on semantic similarity of feature descriptions in a sentence-embedding space, computed before any data value or label is used. Redundancy is therefore handled structurally rather than through pairwise statistics. The clustering is computed once and reused across selection statistics, and relevance is scored only within each cluster rather than across the full feature set.

III. BACKGROUND AND FOUNDATIONS

This section briefly outlines the background and foundations for the presented study: the embedding model used to convert feature descriptions into vectors, the clustering algorithms that organize these vectors into groups, and the label-dependent selection statistics that select a representative feature from each cluster.

- BAAI General Embedding_large variant (BGE_large) Embedding Model - BGE_large is a transformer-based language embedding model that maps a piece of text to a fixed-length dense vector, trained so that semantically similar texts produce vectors with high cosine similarity. This study uses the English checkpoint bge-large-en-v1.5, which maps each text to a 1024-dimensional vector [17]. As recommended for this model, each description is prefixed with the instruction "Represent this sentence for retrieval:" before encoding, and the resulting embeddings are L2-normalized, so the cosine similarity between two embeddings a and b , as shown in Eq. (1), reduces to a dot product.

$$\cos(\theta) = \frac{a \cdot b}{\|a\| \|b\|} \quad (1)$$

The introduction of the Massive Text Embedding Benchmark (MTEB) established a standard multi-task benchmark for comparing embedding models [16], with BGE emerging as a top-performing embedding model and maintaining competitive performance [17,18].

- Clustering Algorithms - Clustering is an established approach often employed for feature engineering. The following are two diverse clustering algorithms, each representing a different established approach to grouping data by similarity.
 - Agglomerative clustering, a bottom-up hierarchical method that merges the two closest clusters at each step, by a chosen linkage criterion, until a target number of clusters is reached [19].
 - DBSCAN clustering, a density-based algorithm that grows clusters wherever points are densely packed and marks sparse points as noise, controlled by two parameters: a distance threshold and a minimum neighborhood size [20].

- Label-dependent Selection Statistics - Three conventional label-dependent statistics are considered for this purpose:

- Random Forest Importance (RFI) scores a feature by its average contribution to impurity reduction across all splits in a Random Forest of T trees, as per Eq. (2) [2,21].

$$\text{Imp}(X_j) = \frac{1}{T} \sum_{t=1}^T \sum_{s \in S(X_j, t)} \frac{N_s}{N} \Delta \text{Gini}(s) \quad (2)$$

where, $S(X_j, t)$ is the set of nodes in the tree t that split on feature X_j , N_s is the number of samples reaching node s , N is the total sample count, and $\Delta \text{Gini}(s)$ is the impurity reduction from that split.

ANOVA F -statistic (ANOVA-F) measures the ratio of between-class to within-class variance for a feature, evaluating each feature independently of the others, as shown in Eq. (3) [22].

$$F = \frac{\sum_{k=1}^K n_k (\bar{x}_k - \bar{x})^2 / (K-1)}{\sum_{k=1}^K \sum_{i=1}^{n_k} (x_{ik} - \bar{x}_k)^2 / (N-K)} \quad (3)$$

where, K is the number of classes, n_k is the sample count in class k , $\bar{x}_k$ the feature's mean within class k , and $\bar{x}$ the overall mean.

- Mutual Information (MI) measures the reduction in uncertainty about a target variable (Y) gained from a feature (X), without assuming a linear relationship between them, as per Eq. (4) [23].

$$I(X; Y) = \sum_x \sum_y p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \quad (4)$$

IV. PROPOSED LLM-EMBEDDING-BASED SEMANTIC FEATURE CLUSTERING PIPELINE

This section presents the workflow (Fig. 1) of the proposed *LLM-embedding-based Semantic Feature Clustering* (LLM-SFC) pipeline for label-dependent feature selection, right from raw feature set to evaluated classifier performance.

1) Level 1: label-agnostic (semantic clustering) – Initially, each feature is described by a single label-agnostic sentence, written from its dataset documentation using a fixed template for each type of feature, or from the attribute name (examples in Table I). This description is then transformed into a numerical representation using a pretrained sentence-embedding language model, BGE_large, producing one dense vector per feature. Thereafter, features are grouped by semantic similarity, i.e., features that end up near each other in the embedded space are grouped using a clustering algorithm. The embeddings are clustered twice, independently, using Agglomerative and DBSCAN clustering algorithms.

The number of clusters k is chosen by maximizing the silhouette score of K -Means over $k = 6$ to $\min(n - 1, 150)$, first in steps of five and then one by one around the best value; the chosen k is then used for agglomerative clustering. Agglomerative clustering uses Ward linkage, which at each step merges the pair of clusters whose union least increases the within-cluster variance, as per Eq. (5).

$$\Delta(A, B) = \frac{|A||B|}{|A|+|B|} \| \mu_A - \mu_B \|^2 \quad (5)$$

TABLE I. FEATURE DESCRIPTION EXAMPLES

Raw Feature	Description
word_freq_internet	Frequency of the word 'internet' in the email
sensor10_transient_increasing_ema_alpha1	Increasing transient exponential moving average response for gas sensor 10
agePct65up	agepercent 65up

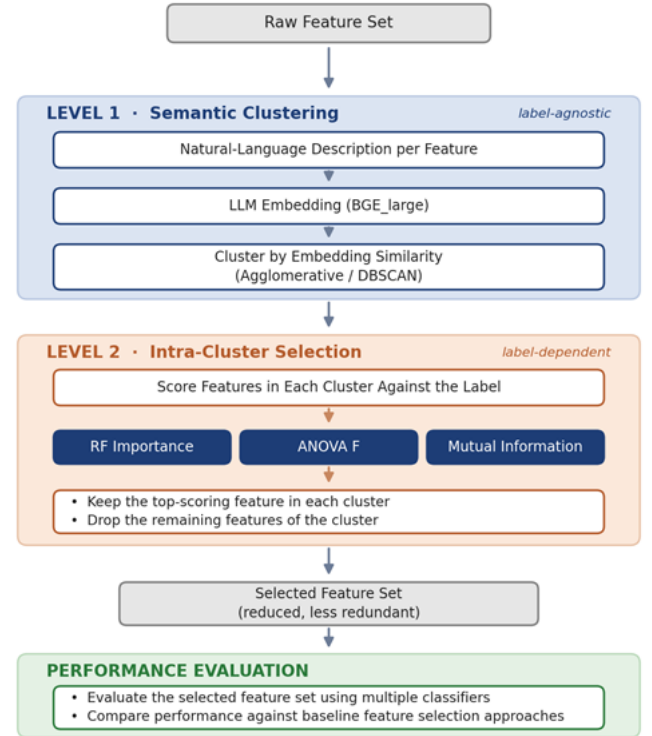


Fig. 1. Proposed LLM-SFC workflow.

where, A and B are two clusters with centroids μ_A and μ_B . Since the embeddings are L2-normalized, $\|a - b\|^2 = 2(1 - \cos \theta)$, so the Euclidean distance used here is a monotone function of the cosine similarity in Eq. (1).

DBSCAN, in contrast, derives its number of clusters from two density parameters: a distance threshold ϵ and a minimum neighborhood size, min_samples . For a point p , its ϵ -neighborhood is defined as per Eq. (6).

$$N_\epsilon(p) = \{q \in E \mid \|p - q\| \leq \epsilon\} \quad (6)$$

where, E is the set of feature embeddings, and ϵ is set to the 90th percentile of each point's distance to its min_samples th nearest neighbor. Clusters grow by connecting core points within ϵ of each other. A point is treated as a core point once its ϵ -neighborhood contains at least min_samples other points.

No label is used at this level. Hence, clustering depends only on what each feature is, never on how it relates to the prediction label, thus allowing a single clustering pass to be reused across every downstream statistic rather than re-clustering for each one.

2) *Level 2: label-dependent* (intra-cluster selection) - In the second level, from within each resulting cluster, one representative feature is chosen as a conventional label-dependent statistic. The proposed pipeline includes three diverse label-dependent selection statistics, RFI, ANOVA-F, and MI, applied only within that cluster rather than across the full feature set. Every feature in a cluster is scored against the true label using whichever statistic is active for that run, and the top scorer is selected while the rest of the cluster is discarded. This step runs independently for each of the three label-dependent selection statistics considered in the proposed LLM-SFC pipeline, on each of the two clusterings, obtained in level one, thus producing six candidate reduced feature sets.

The obtained candidate reduced feature sets are then provided as input to multiple classifiers, under identical cross-validation folds, so that results are directly comparable. Further, each reduced feature subset's performance is compared against the original unreduced feature set baseline, using a Wilcoxon signed-rank test, paired across folds and corrected for multiple comparisons. Algorithm 1 depicts the corresponding procedure.

Algorithm 1. LLM-SFC pipeline for label-dependent feature selection

Input: Feature set $F = \{f_1, \dots, f_n\}$, data matrix X , labels y ,
clustering method M in $\{\text{Agglomerative}, \text{DBSCAN}\}$,
selection statistic S in $\{\text{RFI}, \text{ANOVA-F}, \text{MI}\}$
Output: Reduced feature subset F' , evaluation results R
// Level 1 - semantic clustering. Labels y are not used here.
1: for each f in F do
2: $d_f \leftarrow$ description of f from dataset documentation
3: $e_f \leftarrow \text{BGE_large}(\text{prefix} + d_f)$ // 1024-d embedding, L2-normalized
4: end for
5: $E \leftarrow \{e_1, \dots, e_n\}$
6: if M is Agglomerative then
7: $k \leftarrow \text{argmax}_k \text{silhouette}(\text{KMeans}(E, k))$, k in $[6, \min(n-1, 150)]$
8: $C \leftarrow \text{Agglomerative}(E, k, \text{ward linkage})$
9: else
10: $\text{min_samples} \leftarrow 1$; $\text{eps} \leftarrow 90\text{th percentile of nearest-neighbor distance}$
11: $C \leftarrow \text{DBSCAN}(E, \text{eps}, \text{min_samples})$
12: end if
// $C = \{C_1, \dots, C_m\}$ groups the n features into m clusters

// Level 2 - one representative per cluster. Labels y are used here.
13: $F' \leftarrow \{\}$
14: for each cluster C_j in C do
15: $f_{\text{best}} \leftarrow \text{argmax over } f \text{ in } C_j \text{ of } S(X[:, f], y)$ // scored within C_j only
16: $F' \leftarrow F' \cup \{f_{\text{best}}\}$
17: end for

```
// Evaluation against the unreduced feature set
18: for each classifier  $m$  in  $\{\text{RF}, \text{SVM}, \text{XGBoost}, \text{NB}, \text{LDA}, \text{LR}, \text{RC}\}$  do
19:    $\text{acc\_all} \leftarrow \text{CrossValidate}(m, X[:, F], y)$  // 10 folds, fixed seed
20:    $\text{acc\_red} \leftarrow \text{CrossValidate}(m, X[:, F'], y)$  // identical folds
21:    $p\_m \leftarrow \text{Wilcoxon}(\text{acc\_red}, \text{acc\_all})$  // paired over folds
22: end for
23: apply Benjamini-Hochberg correction within each dataset (21 tests),  $\alpha = 0.05$ 
24: return  $F'$ ,  $R = \{(m, \text{acc\_all}, \text{acc\_red}, p\_m)\}$ 
```

V. RESULTS AND DISCUSSION

The candidate reduced feature sets generated through the proposed LLM-SFC pipeline are provided as input to seven classifiers, namely RF, Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), Naive Bayes (NB), Linear Discriminant Analysis (LDA), Logistic Regression (LR), and Ridge Classifier (RC), under identical cross-validation folds. Performance results for each of these seven classifiers are reported for each of the three label-dependent selection statistics (i.e., RFI, ANOVA-F, and MI), across both clustering methods (Agglomerative and DBSCAN) considered in this study.

A. Dataset Description

In order to assess and establish the generalization capability of the proposed LLM-SFC pipeline, performance evaluation was done across heterogeneous data domains using three dissimilar and diverse experimental datasets, so that any pattern holding across all three is unlikely to be specific to a single domain. The following three experimental datasets were used in this study:

1) *Dataset_Gas Sensor Drift (DS_GSD)* contains 13,910 readings from an array of 16 metal-oxide gas sensors, described by 128 steady-state and transient response features [24]. A stratified sample of 5,000 readings was used for all experiments, with the task of identifying which of six gases the array was exposed to.

2) *Dataset_Spambase (DS_SPB)* contains 4,601 emails described by 57 features capturing word frequencies, character frequencies, and capitalization patterns, with the task of classifying each email as spam or legitimate [25].

3) *Dataset_Communities and Crime (DS_CAC)* contains 1,994 communities described by 127 socioeconomic and law-enforcement attributes, with the task of classifying each community's violent crime rate as high or low relative to a threshold [26].

B. Experimental Setup

Missing values are replaced by the feature median, and all features are scaled to $[0, 1]$ using Min-Max scaling. Every feature set is evaluated with stratified 10-fold cross-validation (shuffled, seed 42), using the same folds for all methods so that results can be compared directly. Seven classifiers are used, with the settings listed in Table II.

TABLE II. EXPERIMENTAL SETTINGS

Component	Setting
Feature descriptions	One template per feature type, from dataset documentation (examples in Table I)
Embedding model	bge-large-en-v1.5; 1024 dimensions; retrieval prefix; L2-normalized
Agglomerative clustering	Ward linkage; k chosen by K-Means silhouette over $k = 6$ to $\min(n - 1, 150)$
DBSCAN	$\min_samples \in \{1, 5, 10\}$; $\epsilon = 90$ th percentile of distance to the $\min_samples$ -th neighbor
Level 2 statistics	RFI (50 trees, per cluster); ANOVA-F ($f_classif$); MI (nearest-neighbor estimator)
Comparators	SelectKBest-MI and mRMR at the same feature count
Preprocessing	Median imputation; Min-Max scaling to $[0, 1]$
Evaluation	Stratified 10-fold CV, shuffled, seed 42; Gas Sensor Drift: stratified sample of 5,000 readings
Random Forest	100 trees
SVM	RBF kernel, $C = 1$, $\gamma = 'scale'$
XGBoost	Library defaults; log-loss evaluation metric
Naive Bayes	Gaussian; default variance smoothing
LDA	SVD solver
Logistic Regression	L2 penalty, $C = 1$, up to 2,000 iterations
Ridge Classifier	$\alpha = 1.0$

Results are reported as mean accuracy with its standard deviation across folds. SelectKBest-MI and mRMR serve as comparators and are always given the same number of features as the LLM-SFC subset they are compared with. The pipeline is implemented in Python using scikit-learn, sentence-transformers, XGBoost, and mrmr-selection, and all other settings are summarized in Table II.

For DBSCAN, values of 1, 5, and 10 were tried for $\min_samples$. The larger two merged the features into only two to four clusters on Gas Sensor Drift and Spambase, so $\min_samples = 1$ was used. With this setting, every point is a core point, and DBSCAN behaves as single-linkage clustering cut at distance ϵ , which gives a useful contrast to the Ward linkage used by agglomerative clustering.

C. Experimental Results

The following subsections report accuracy obtained for each classifier under consideration, for each experimental dataset under both clustering methods considered in this study. Additionally, performance results were also computed for the original unreduced feature set (*All_Features*), along with reduced feature sets generated using SelectKBest with MI (*SelectKBest-MI*) and mRMR feature selection strategies. These three are used as baselines for comparison. All results are reported as the average accuracy across the cross-validation folds, together with the standard deviation, to indicate the consistency of each outcome.

TABLE III. PERFORMANCE COMPARISON OF BASELINE AND PROPOSED LLM-SFC FEATURE SELECTION STRATEGIES ON GAS SENSOR DRIFT DATASET

Classifier				Agglomerative Clustering			DBSCAN Clustering		
Feature Selection (Count)	All_Features (128)	SelectKBest-MI (58)	mRMR (58)	RFI (58)	ANOVA-F (58)	MI (58)	RFI (18)	ANOVA-F (18)	MI (18)
RF	97.12±0.46	98.26±0.57	98.48±0.45	98.98±0.35	99.24±0.34	99.14±0.37	97.96±0.46	97.86±0.51	98.08±0.62
SVM	82.82±1.21	78.40±1.49	80.02±1.48	85.32±1.22	85.18±1.29	85.22±1.32	79.54±2.03	82.80±1.83	78.40±1.60
XGBoost	96.92±0.55	98.48±0.50	98.64±0.61	98.86±0.47	98.94±0.51	98.94±0.44	97.78±0.72	97.96±0.79	98.18±0.66
NB	51.94±1.16	53.78±1.83	52.22±1.81	50.66±1.07	55.88±1.44	55.50±1.18	45.06±1.42	42.54±1.70	41.20±1.58
LDA	90.94±1.11	91.56±1.51	92.01±1.51	93.26±1.07	92.66±1.34	93.40±1.16	75.44±2.53	79.48±2.32	78.88±2.22
LR	90.02±0.83	85.34±1.39	85.98±1.15	90.74±0.96	90.64±1.02	90.48±1.07	78.02±1.64	78.26±1.82	72.86±1.71
RC	90.80±1.30	83.12±0.64	86.12±0.64	91.34±1.50	92.78±1.29	92.64±1.37	69.60±1.58	75.18±1.46	72.74±1.78

Table III presents the results for comparison of classifier performance under baseline and proposed LLM-SFC feature selection strategies for the Gas Sensor Drift dataset, which has 128 features. Agglomerative clustering targets $k = 58$ clusters (54.7% reduction), while DBSCAN, with $\min_samples = 1$,

converges on 18 features (85.9% reduction). The baseline feature sets reach a mean accuracy of 85.79% (*All_Features*), 84.13% (*SelectKBest-MI*), and 84.78% (*mRMR*), while the agglomerative-reduced feature set achieves a mean accuracy of 87.02% under RFI and 87.90% under both ANOVA-F and MI.

RF, XGBoost, and SVM each gain roughly two points under every strategy, and LDA, LR, and RC gain between half a point and two and a half points under agglomerative clustering. NB is the only classifier that does not benefit uniformly under agglomerative clustering and falls 1.28 points below baseline under RFI, though gaining 3.94 and 3.56 points under ANOVA-F and MI, respectively. DBSCAN's much smaller, density-driven reduction to 18 features costs considerably more, with accuracy dropping well below baseline across most classifiers and strategies.

Table IV presents the results for comparison of classifier performance under baseline and proposed LLM-SFC feature selection strategies for the Spambase dataset, which has 57 features. Agglomerative clustering targets $k = 44$ clusters (22.8% cut), while DBSCAN, with $\text{min_samples} = 1$,

converges on just 8 features (86% cut), reflecting a very aggressive reduction. The baseline feature sets reach a mean accuracy of 88.58% (All_Features), 86.49% (SelectKBest-MI), and 88.90% (mRMR), while the agglomerative-reduced set achieves 89.36% under RFI, 89.47% under ANOVA-F, and 89.54% under MI, an improvement under every strategy, though not uniformly across classifiers. Under agglomerative clustering, RF, SVM, LDA, and RC gain between one and two and a half points, XGBoost gains under a point, and LR is effectively unchanged. NB again runs against the trend, losing between 1.42 and 1.96 points. DBSCAN's aggressive reduction to 8 features costs considerably more broadly, with every classifier losing meaningful ground, including RF and XGBoost, which were essentially untouched under agglomerative.

TABLE IV. PERFORMANCE COMPARISON OF BASELINE AND PROPOSED LLM-SFC FEATURE SELECTION STRATEGIES ON SPAMBASE DATASET

Classifier				Agglomerative Clustering			DBSCAN Clustering		
Feature Selection (Count)	All Features (57)	SelectKBest-MI (44)	mRMR (44)	RFI (44)	ANOVA-F (44)	MI (44)	RFI (8)	ANOVA-F (8)	MI (8)
RF	93.46±0.66	92.20±0.85	94.18±0.66	95.07±0.59	95.59±0.63	95.65±0.57	87.85±1.76	85.57±1.27	87.29±1.63
SVM	91.31±0.71	89.98±1.17	91.87±0.85	92.81±0.71	93.24±0.72	93.18±0.74	81.16±1.48	79.87±1.92	80.92±1.66
XGBoost	94.39±0.76	91.87±0.80	94.85±0.95	95.07±1.05	95.26±1.04	95.18±1.00	86.79±1.30	85.39±1.19	87.92±1.33
NB	80.66±1.48	75.35±1.56	80.28±1.07	79.18±1.35	78.70±1.25	79.24±1.20	67.92±0.98	71.44±1.53	69.98±1.05
LDA	85.72±0.95	85.07±1.04	86.18±1.28	87.66±1.31	87.85±1.41	87.74±1.48	72.53±1.70	76.18±1.44	72.42±1.83
LR	88.24±1.08	86.22±0.85	87.96±0.80	88.29±1.24	88.20±1.05	88.35±1.26	70.22±1.51	75.98±1.07	70.22±1.36
RC	86.29±0.98	84.76±1.11	87.02±1.32	87.44±1.55	87.46±1.35	87.44±1.51	71.31±1.75	76.03±1.44	71.25±1.87

TABLE V. PERFORMANCE COMPARISON OF BASELINE AND PROPOSED LLM-SFC FEATURE SELECTION STRATEGIES ON COMMUNITIES AND CRIME DATASET

Classifier				Agglomerative Clustering			DBSCAN Clustering		
Feature Selection (Count)	All Features (127)	SelectKBest-MI (56)	mRMR (56)	RFI (56)	ANOVA-F (56)	MI (56)	RFI (22)	ANOVA-F (22)	MI (22)
RF	84.36±1.64	85.46±1.53	86.31±1.12	86.31±1.96	86.61±1.91	86.36±1.92	85.81±3.96	86.61±3.46	86.46±3.65
SVM	84.81±1.45	84.95±1.39	85.10±2.91	86.01±1.41	85.76±1.68	85.91±1.21	86.16±3.07	86.01±2.94	85.35±3.38
XGBoost	85.11±1.11	84.85±1.13	85.25±1.27	86.21±1.60	85.71±1.39	85.81±1.13	85.00±4.13	85.36±4.17	85.31±3.37
NB	80.34±1.00	79.74±1.53	81.25±1.12	81.79±1.02	82.39±1.08	82.44±0.68	82.85±3.19	82.04±3.63	81.94±3.78
LDA	83.36±1.41	84.65±2.41	85.30±1.10	85.50±1.31	85.81±0.89	85.40±1.77	85.75±3.55	85.86±3.34	85.91±3.20
LR	85.61±1.21	84.85±1.97	85.80±1.98	86.16±1.25	85.70±1.17	85.81±0.63	85.45±2.91	86.01±2.86	85.86±3.04
RC	84.61±1.09	84.60±1.03	85.10±1.14	85.60±1.50	85.40±1.21	85.00±1.31	85.15±3.36	85.66±3.31	85.96±2.83

Table V presents the results for comparison of classifier performance under baseline and proposed LLM-SFC feature selection strategies for the Communities and Crime dataset, which has 127 features. Agglomerative clustering targets $k = 56$ clusters (55.9% reduction), whereas DBSCAN, with $\text{min_samples} = 1$, converges on 22 features (82.7% cut). The baseline feature sets reach a mean accuracy of 84.03% (All_Features), 84.16% (SelectKBest-MI), and 84.87% (mRMR), while the agglomerative-reduced set achieves 85.37%, 85.34%, and 85.25% under RFI, ANOVA-F, and MI, respectively, exceeding the baseline in every case. For this dataset, the reduction is most uniformly beneficial under

agglomerative clustering, as all combinations land above the baseline, with gains ranging from 0.09 to 2.45 points. Moreover, unlike on DS_GSD and DS_SPB, for DS_CAC, the DBSCAN-reduced feature set performs comparably to the agglomerative-reduced feature set across every classifier.

Fig. 2 presents a comparative visual of the mean accuracy attained by the agglomerative-reduced feature set for the three label-dependent selection statistics considered in this study, relative to the mean accuracy attained by the unreduced baseline, SelectKBest-MI and mRMR feature sets, for the three experimental datasets.

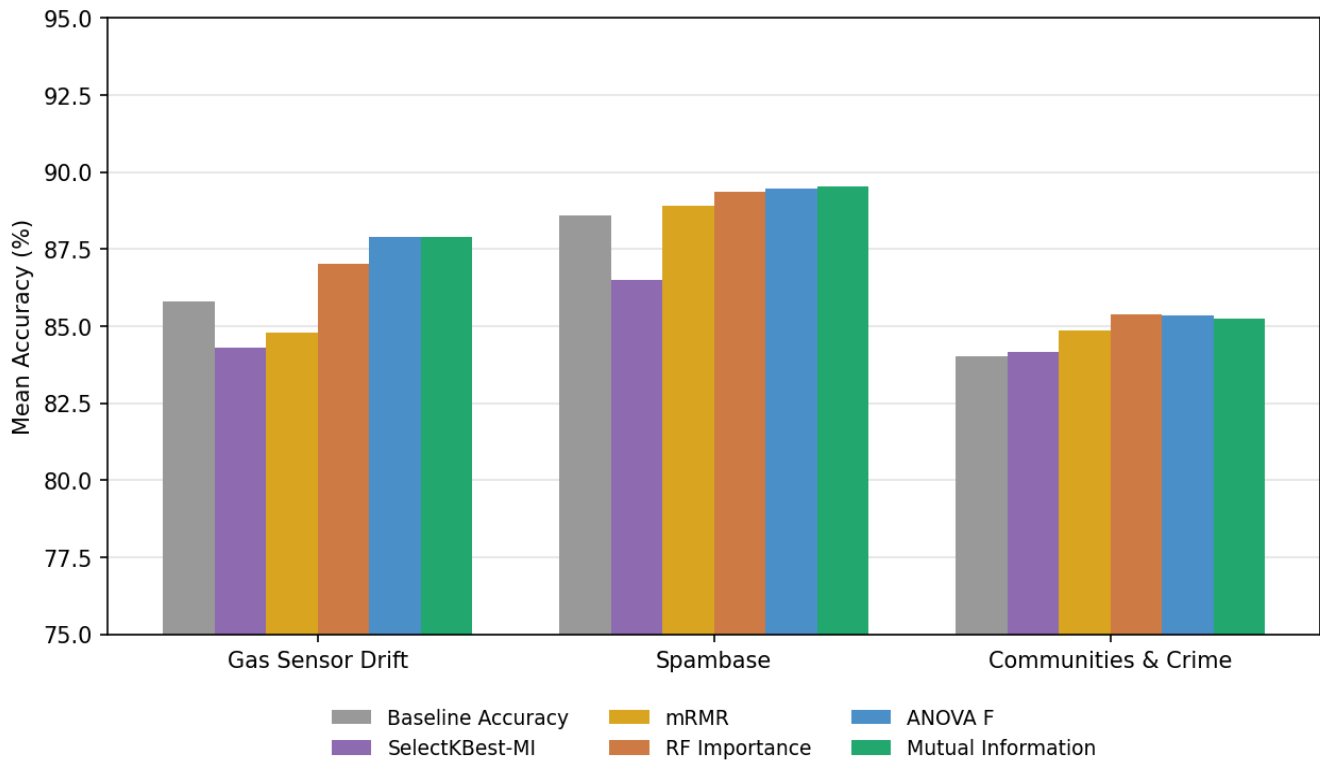


Fig. 2. Mean accuracy comparison across experimental datasets.

VI. SIGNIFICANCE TESTING

A small difference in mean accuracy does not necessarily indicate that one selection statistic (RFI, ANOVA-F or MI) is better or worse than another, as such differences can ensue due to normal variation across runs. To check whether the differences are statistically meaningful, the Wilcoxon signed-rank test was performed to compare each selection statistic's accuracy against the baseline feature set for every classifier, paired across cross-validation folds. As a large number of comparisons were involved, Benjamini-Hochberg False Discovery Rate correction was applied within each dataset before any result was classified as significant, with significance assessed at $\alpha = 0.05$ on the corrected p -values. Based on the

obtained accuracy results (presented in Section V), it can be observed that agglomerative-reduced feature sets performed better than the baseline feature set for all three experimental datasets, and so significance testing results are reported for them only. The tables below present the significance test results, showing the percentage difference in accuracy (Δ) between the baseline and reduced feature sets after feature selection. Statistically significant outcomes are marked in bold font.

Table VI presents the significance test results for the Gas Sensor Drift dataset. It can be seen that 20 of 21 combinations fare better than the baseline feature set, with gains between 0.46 and 3.94 points, and 16 of those gains are statistically significant.

TABLE VI. SIGNIFICANCE TEST RESULTS FOR GAS SENSOR DRIFT DATASET

Classifier	RFI			ANOVA-F			MI		
	Δ (%)	p -value	Significant	Δ (%)	p -value	Significant	Δ (%)	p -value	Significant
RF	+1.86	0.004	Yes	+2.12	0.003	Yes	+2.02	0.003	Yes
SVM	+2.50	0.003	Yes	+2.36	0.003	Yes	+2.40	0.003	Yes
XGBoost	+1.94	0.004	Yes	+2.02	0.003	Yes	+2.02	0.003	Yes
NB	-1.28	0.24	No	+3.94	0.002	Yes	+3.56	0.002	Yes
LDA	+2.32	0.003	Yes	+1.72	0.004	Yes	+2.46	0.003	Yes
LR	+0.72	0.084	No	+0.62	0.059	No	+0.46	0.072	No
RC	+0.54	0.065	No	+1.98	0.004	Yes	+1.84	0.004	Yes

Table VII presents the significance test results for the Spambase dataset. It can be seen that 17 of 21 combinations fare

better than the baseline feature set, and 12 of those gains are statistically significant.

TABLE VII. SIGNIFICANCE TEST RESULTS FOR SPAMBASE DATASET

Classifier	RFI			ANOVA-F			MI		
	Δ (%)	<i>p-value</i>	Significant	Δ (%)	<i>p-value</i>	Significant	Δ (%)	<i>p-value</i>	Significant
<i>RF</i>	+1.61	0.004	Yes	+2.13	0.003	Yes	+2.19	0.003	Yes
<i>SVM</i>	+1.50	0.004	Yes	+1.93	0.004	Yes	+1.87	0.004	Yes
<i>XGBoost</i>	+0.68	0.073	No	+0.87	0.077	No	+0.79	0.085	No
<i>NB</i>	-1.48	0.3	No	-1.96	0.39	No	-1.42	0.48	No
<i>LDA</i>	+1.94	0.004	Yes	+2.13	0.003	Yes	+2.02	0.004	Yes
<i>LR</i>	+0.05	1.0	No	-0.04	1.0	No	+0.11	0.50	No
<i>RC</i>	+1.15	0.005	Yes	+1.17	0.005	Yes	+1.15	0.005	Yes

Table VIII presents the significance test results for the Communities and Crime dataset. It can be seen that all 21 combinations land above the baseline feature set, with gains between 0.09 and 2.45 points, and 12 of those gains being

significant. For this dataset, with over half the features removed, accuracy improves rather than merely being preserved, with no significant loss anywhere.

TABLE VIII. SIGNIFICANCE TEST RESULTS FOR COMMUNITIES AND CRIME DATASET

Classifier	RFI			ANOVA-F			MI		
	Δ (%)	<i>p-value</i>	Significant	Δ (%)	<i>p-value</i>	Significant	Δ (%)	<i>p-value</i>	Significant
<i>RF</i>	+1.95	0.004	Yes	+2.25	0.003	Yes	+2.00	0.004	Yes
<i>SVM</i>	+1.20	0.005	Yes	+0.95	0.082	No	+1.10	0.005	Yes
<i>XGBoost</i>	+1.10	0.005	Yes	+0.60	0.079	No	+0.70	0.054	No
<i>NB</i>	+1.45	0.004	Yes	+2.05	0.004	Yes	+2.10	0.004	Yes
<i>LDA</i>	+2.14	0.003	Yes	+2.45	0.003	Yes	+2.04	0.004	Yes
<i>LR</i>	+0.55	0.061	No	+0.09	1.0	No	+0.20	0.66	No
<i>RC</i>	+0.99	0.10	No	+0.79	0.067	No	+0.39	0.72	No

VII. CONCLUSION

Existing feature selection approaches often evaluate feature relevance using individual scoring criteria, sequentially followed by a separate selection step. Though some unsupervised approaches, such as clustering, might consider the underlying relationships among features, relevance and redundancy are not always jointly modeled within a combined mechanism. Consequently, redundant features may receive high relevance scores and get considered as candidates for selection, potentially reducing the discriminative quality of the resulting feature subset.

This study presented LLM-embedding-based Semantic Feature Clustering (LLM-SFC), a two-level pipeline that clusters features by the semantic content of their descriptions before a conventional label-dependent statistic selects a representative feature from each cluster. In the first level, which is label-agnostic, a natural-language description is written for each feature, conveying its meaning independently of the label. These descriptions are then converted into numerical representations using a pretrained sentence embedding model. The resulting embeddings are clustered according to semantic similarity. In the second level, which is label-dependent, a representative feature is selected from each cluster formed in the previous level, using a label-dependent statistical criterion, thereby reducing redundancy while retaining features most relevant to the label. The proposed LLM-SFC pipeline employs

two varied clustering algorithms, Agglomerative and DBSCAN, for semantic clustering. Further, three prevalent label-dependent selection statistics, namely RFI, ANOVA-F, and MI, are used for selecting the representative feature.

Across three datasets from unrelated domains, agglomerative clustering reduced the feature set by 22.8–55.9% while improving mean accuracy in every case, with 40 of 63 classifier-statistic combinations reaching statistical significance and no significant losses anywhere. DBSCAN's reduction ratio depends on data density, and it led to aggressive reduction on two of the experimental datasets and hence to much lower accuracy. On the third experimental dataset, where the reduction is more moderate, its performance was found to be comparable with that of agglomerative clustering. As agglomerative-reduced feature sets fared better than the baseline feature set, significance testing was also performed for these feature sets, for all three experimental datasets. No dataset was found to register a single significant loss. What varied between datasets was only how many of the gains cleared the significance threshold. Out of 21 combinations, 16 significant gains were found for the Gas Sensor Drift dataset, and 12 for both the Spambase dataset and Communities and Crime dataset.

A limitation of the presented work is that feature descriptions were authored for this study rather than drawn from an independently validated source, so description quality is a variable the clustering process inherits. Replacing the label-

dependent statistic with a label-free one over the same clustering structure is a natural next step, opening this pipeline to settings where labels are costly, delayed, or unavailable.

REFERENCES

- [1] D. Theng and K. K. Bhoyar, "Feature selection techniques for machine learning: A survey of more than two decades of research," *Knowledge and Information Systems*, vol. 66, no. 13, 2023, doi: 10.1007/s10115-023-02010-5.
- [2] M. Buyukkececi and M. C. Okur, "A comprehensive review of feature selection and feature selection stability in machine learning," *Gazi University Journal of Science*, vol. 36, no. 4, pp. 1506–1520, 2023, doi: 10.35378/gujs.993763.
- [3] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
- [4] S. Xie, Y. Zhang, D. Lv, X. Chen, J. Lu, and J. Liu, "A new improved maximal relevance and minimal redundancy method based on feature subset," *The Journal of Supercomputing*, vol. 79, pp. 3157–3180, 2023, doi: 10.1007/s11227-022-04763-2.
- [5] P. Bhattacharjee and P. Mitra, "A survey of density-based clustering algorithms," *Frontiers of Computer Science*, vol. 15, no. 1, Art. no. 151308, 2021, doi: 10.1007/s11704-019-9059-3.
- [6] G. J. Oyewole and G. A. Thopil, "Data clustering: Application and trends," *Artificial Intelligence Review*, vol. 56, no. 7, pp. 6439–6475, 2023, doi: 10.1007/s10462-022-10325-y.
- [7] S. Seth, S. Mallik, T. Bhadra, and Z. Zhao, "Dimensionality reduction and Louvain agglomerative hierarchical clustering for cluster-specified frequent biomarker discovery in single-cell sequencing data," *Frontiers in Genetics*, vol. 13, Art. no. 828479, 2022, doi: 10.3389/fgene.2022.828479.
- [8] S.-G. Fang, D. Huang, C.-D. Wang, and Y. Tang, "Joint multi-view unsupervised feature selection and graph learning," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 8, no. 1, pp. 16–31, 2024, doi: 10.1109/TETCI.2023.3306233.
- [9] S. Zhou, P. Song, Z. Song, and L. Ji, "Soft-label guided non-negative matrix factorization for unsupervised feature selection," *Expert Systems with Applications*, vol. 216, Art. no. 119468, 2023, doi: 10.1016/j.eswa.2022.119468.
- [10] Z. Li, F. Nie, D. Wu, Z. Hu, and X. Li, "Unsupervised feature selection with weighted and projected adaptive neighbors," *IEEE Transactions on Cybernetics*, vol. 53, no. 2, pp. 1260–1271, 2023, doi: 10.1109/TCYB.2021.3087632.
- [11] Z. Wang, Q. Li, H. Zhao, and F. Nie, "Simultaneous local clustering and unsupervised feature selection via strong space constraint," *Engineering Applications of Artificial Intelligence*, vol. 161, Art. no. 112121, 2025.
- [12] D. P. Jeong, Z. C. Lipton, and P. Ravikumar, "LLM-Select: Feature selection with large language models," *Transactions on Machine Learning Research*, 2025.
- [13] S. Hegselmann, A. Buendia, H. Lang, M. Agrawal, X. Jiang, and D. Sontag, "TabLLM: Few-shot classification of tabular data with large language models," in *Proc. 26th Int. Conf. Artificial Intelligence and Statistics (AISTATS)*, vol. 206, pp. 5549–5581, 2023.
- [14] Y. Gu, H. You, J. Cao, M. Yu, H. Fan, and S. Qian, "Large language models for constructing and optimizing machine learning workflows: A survey," *arXiv preprint arXiv:2411.10478*, 2024.
- [15] B. Koloski, A. Margeloiu, X. Jiang, B. Škrlić, N. Simidjievski, and M. Jamnik, "LLM embeddings for deep learning on tabular data," *arXiv preprint arXiv:2502.11596*, 2025.
- [16] N. Muennighoff, N. Tazi, L. Magne, and N. Reimers, "MTEB: Massive text embedding benchmark," in *Proc. 17th Conf. European Chapter of the Association for Computational Linguistics (EACL)*, pp. 2014–2037, 2023, doi: 10.18653/v1/2023.eacl-main.148.
- [17] S. Xiao, Z. Liu, P. Zhang, N. Muennighoff, D. Lian, and J.-Y. Nie, "C-Pack: Packed resources for general Chinese embeddings," in *Proc. 47th Int. ACM SIGIR Conf. Research and Development in Information Retrieval*, pp. 641–649, 2024, doi: 10.1145/3626772.3657878.
- [18] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu, "BGE M3-Embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation," *arXiv preprint arXiv:2402.03216*, 2024.
- [19] L. M. C. Cabezas, R. Izbecki, and R. B. Stern, "Hierarchical clustering: Visualization, feature importance and model selection," *Applied Soft Computing*, vol. 141, Art. no. 110303, 2023, doi: 10.1016/j.asoc.2023.110303.
- [20] S. Chowdhury, N. Helian, and R. Cordeiro de Amorim, "Feature weighting in DBSCAN using reverse nearest neighbours," *Pattern Recognition*, vol. 137, Art. no. 109314, 2023, doi: 10.1016/j.patcog.2023.109314.
- [21] M. Rotari and M. Kulahci, "Variable selection wrapper in presence of correlated input variables for random forest models," *Quality and Reliability Engineering International*, vol. 40, no. 1, pp. 297–312, 2024, doi: 10.1002/qre.3398.
- [22] F. Golnoori, F. Zamani Boroujeni, and S. A. Monadjemi, "A comparative study on deep feature selection methods for skin lesion classification," *IET Image Processing*, vol. 18, no. 4, pp. 996–1013, 2024, doi: 10.1049/ipr2.13002.
- [23] H. Gong, Y. Li, J. Zhang, B. Zhang, and X. Wang, "A new filter feature selection algorithm for classification task by ensembling Pearson correlation coefficient and mutual information," *Engineering Applications of Artificial Intelligence*, vol. 131, Art. no. 107865, 2024, doi: 10.1016/j.engappai.2024.107865.
- [24] A. Vergara, "Gas Sensor Array Drift Dataset," *UCI Machine Learning Repository*, 2012, doi: 10.24432/C5RP6W.
- [25] M. Hopkins, E. Reeber, G. Forman, and J. Suermondt, "Spambase," *UCI Machine Learning Repository*, 1999, doi: 10.24432/C53G6X.
- [26] M. Redmond, "Communities and Crime," *UCI Machine Learning Repository*, 2002, doi: 10.24432/C53W3X.
- [27] 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, doi: 10.1109/TPAMI.2005.159.
- [28] Q. Song, J. Ni, and G. Wang, "A fast clustering-based feature subset selection algorithm for high-dimensional data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 1, pp. 1–14, 2013, doi: 10.1109/TKDE.2011.181.