COMPARATIVE DATA MINING APPROACH FOR DISCOVERING PATIENT RISK GROUPS IN HEALTHCARE DATA

1st Irfanuzzaman Montasir
Dept. of CSE
United International University
Dhaka, Bangladesh
011221276

2nd Md. Khademul Islam Nahin
Dept. of CSE
United International University
Dhaka, Bangladesh
011221282

3rd Ashikur Rahaman
Dept. of CSE
United International University
Dhaka, Bangladesh
011221292

4th Md. Mizanur Rahman
Dept. of CSE
United International University
Dhaka, Bangladesh
011221381

5th Argha Biswas
Dept. of CSE
United International University
Dhaka, Bangladesh
011221293

***Abstract*—Machine learning model performance relies heavily on feature engineering, especially when it comes to medical diagnosis as both accurate predictions and accurate interpretations are needed. The purpose of this research was to examine whether cluster-derived features would enhance the classification accuracy of four different datasets like Diabetes, Heart Disease, Hepatitis and Chronic Kidney Disease that represented a range of disease types. To obtain cluster derived features, we applied multiple dimensionality reduction techniques (Principal Component Analysis, t-SNE, UMAP) to each of the datasets, using multiple clustering algorithms (KMeans, Hierarchical, DBSCAN), and then added these derived features to three different classification models (Logistic Regression, Random Forest, XGBoost). Our results show that the effect of adding cluster derived features will vary depending upon which dataset you use, but generally did not universally increase the accuracy of the models tested. While there were modest increases in accuracy for some datasets such as the Heart Disease and Hepatitis datasets,the accuracy of the other two datasets Kidney Disease and Diabetes which is actually decreased slightly after applying cluster derived features. There was no significant change in accuracy for either the Heart Disease and Diabetes datasets when using tree based models (Random Forest, XGBoost). Overall, our results demonstrate that cluster-derived features have the potential to be beneficial when the following conditions are met: (1) the cluster features identify interaction patterns that were not captured in the original features; (2) the baseline model performance is not at near-perfect levels, and (3) the clusters reflect a good alignment of risk groups of patients with the disease outcome. Although we found no statistically significant differences in predictive accuracy based on the addition of cluster derived features to our machine learning models, the use of clustering has provided clear clinical interpretability to our results through the identification of separate patient risk groups. Ultimately, the results of this study provide empirical evidence regarding under what conditions cluster-derived feature engineering is successful and unsuccessful which is providing an understanding of how specific characteristics of datasets and architectures of machine learning models contribute to the success or failure of feature engineering.**

***Index Terms*—Machine Learning, Feature Engineering, Clustering, Medical Diagnosis, Disease Prediction, Dimensionality Reduction, Patient Stratification**

I. Introduction

Medical disease prediction is increasingly important for preventive healthcare, allowing early intervention and personalized treatment strategies. Machine learning models have been demonstrating great promise in this area using patient data to predict the occurrence of diseases with high accuracy [1]. However, the performance of these models depends heavily upon the quality and informativeness of the input features.

Feature engineering, which is creating new features from existing data, is an essential step in improving model performance. In medical applications, well-crafted features can capture complex relationships between clinical measurements that may reveal patterns invisible to raw data analysis [2]. Clustering algorithms are one method for feature engineering by identifying natural groupings within the data, which can then be encoded as categorical features for supervised learning tasks.

A. Motivation

The motivation behind this study comes from two key observations:

1) **Clinical Interpretability**: Medical practitioners not only want accurate predictions but also need models they can understand why patients were classified as high-risk [3].

2) **Pattern Discovery**: Unsupervised clustering has the potential to identify subgroups of patients characterized by distinct disease manifestations, which could enhance predictive modeling.

B. Research Questions

This study will answer the following research questions:

1) Does including cluster-derived features improve classification accuracy for medical disease prediction?
2) What configurations of clustering (algorithm, dimensionality reduction, number of clusters) produce the most meaningful patient stratification?
3) Do different classification models benefit equally from cluster-based features?

C. Contributions

We believe our primary contributions include:

- A systematic evaluation framework comparing classification performance with and without cluster-derived features across multiple medical datasets
- Empirical evidence suggests that cluster features will not always result in improved accuracy when baseline features are already highly discriminative
- We will show clustering's value for patient risk stratification and clinical interpretability independent of predictive accuracy gains
- Open-source code and methodology for reproducible research in medical ML

II. Related Work

A. The Role of Feature Engineering in Medical Machine Learning

Feature engineering has historically been identified as one of the most important factors of successful machine learning (ML) applications; Domingos stated that "feature engineering is the key" to practical success in ML [4]; and, as such, researchers in medical domains have explored numerous types of feature constructions, such as polynomial features, interaction terms, and domain-specific transformations [16].

More recent studies have also demonstrated the effectiveness of feature construction using risk scores for predicting cardiovascular diseases [17], as well as demonstrated how deep learning can be used to automatically generate hierarchical feature representations from electronic health records (EHRs) [18].

B. Using Clustering for Pattern Discovery

Clustering has been widely utilized in medical data analysis to perform patient stratification and to identify subtypes of disease. For example, Ahmad and Hashmi [19] compared K-means clustering to hierarchical clustering on heart disease data, and Perveen et al. [9] compared clustering algorithms to determine their performance in predicting diabetes.

Li et al. [10] proposed a framework called Patient Similarity Learning that combined clustering with supervised learning for disease prediction, and found that cluster memberships may be useful as an additional source of information when constructing features, which motivated our investigation.

C. Hybrid Methods

There has been considerable interest in combining unsupervised and supervised learning methods. Wagstaff et al. [11] introduced constrained clustering to enable the incorporation of domain knowledge into the clustering process. Additionally, there has recently been research on developing hybrid clustering and feature learning methods. For example, Xie et al. [12] proposed a method called Deep Embedded Clustering that jointly optimized clustering and feature learning.

However, very little research has investigated whether cluster-derived features will consistently improve classification accuracy when applied across multiple medical datasets and architectures, which we address here.

D. Dimensionality Reduction

Dimensionality reduction techniques such as Principal Component Analysis (PCA) [13], t-distributed Stochastic Neighbor Embedding (t-SNE) [14], and Uniform Manifold Approximation and Projection (UMAP) [15] are commonly employed with clustering to handle the high dimensionality of medical data. Specifically, t-SNE was developed by van der Maaten [14] and is now a popular technique for visualizing the underlying structure of large, high-dimensional datasets prior to performing clustering.

III. Methodology

The methodology employed in this research is divided into 4 main stages: (1) Data processing, (2) Clustering-based generation of features, (3) Classification, (4) Comparative assessment of results.

The entire methodology can be seen as a series of parallel processes that run simultaneously and depicted as a whole in Figure 1.

A. Datasets

To assess the proposed method, we used 4 public datasets available on the University of California at Irvine (UCI) and Kaggle.
Machine Learning Repository:

1) **Diabetes Dataset** (768 samples, 9 attributes). The task was to predict the onset of diabetes on the basis of diagnostic indicators.
2) **Heart Disease Dataset** (1,025 samples, 14 attributes). The task was to predict cardiovascular disease.
3) **Hepatitis Dataset**. The task was to predict the outcomes of hepatitis diseases.
4) **Chronic Kidney Disease Dataset** (400 samples). The task was to diagnose chronic kidney disease (CKD) based on clinical measures including serum creatinine, blood urea, blood pressure, hemoglobin, and specific gravity.

All the datasets include both numerical and nominal characteristics and their classification target is binary.

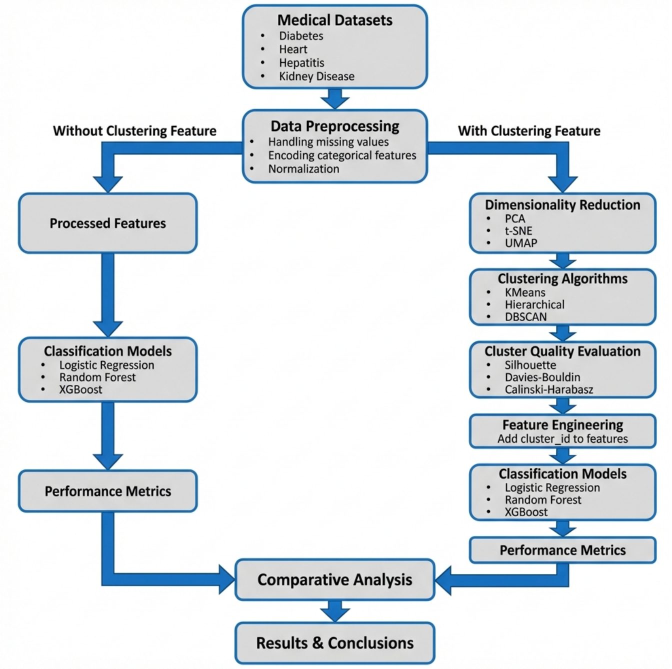

Fig. 1. Overall methodology workflow showing parallel paths with and without cluster features.

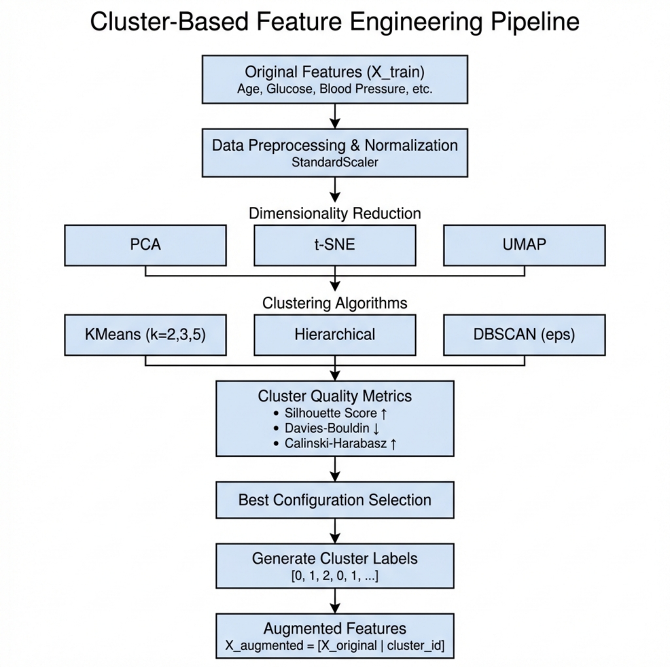

Fig. 2. Detailed cluster-based feature engineering pipeline showing dimensionality reduction, clustering, and feature augmentation steps.

B. Data Preprocessing

Data preprocessing includes the following steps:

1) **Missing value replacement**: For numerical characteristics missing values are replaced by median values; for nominal characteristics mode imputation is applied.
2) **Encoding**: Nominal features are encoded using Label Encoding.
3) **Standardization**: Characteristics are normalized to have zero mean and unit variance using Standard Scaler.
4) **Split of training and test sets**: Stratified 80-20 split of training set and test set to preserve balance among classes.

C. Cluster-Based Feature Engineering

Figure 2 details the clustering pipeline for feature generation.

1) Dimensionality Reduction: We apply three dimensionality reduction techniques:

- **PCA**: Captures maximum variance in 2-5 components
- **t-SNE**: Preserves local structure, perplexity=30, n_components=2
- **UMAP**: Balances local and global structure preservation, n_neighbors=15

2) Clustering Algorithms: Three clustering algorithms are evaluated:

- **KMeans**: Tested with k $\in \{2, 3, 5\}$, random_state=42
- **Hierarchical Agglomerative Clustering**: Ward linkage criterion
- **DBSCAN**: eps $\in \{0.3{,}0.5{,}0.7{,}0.9{,}1.1\}$, min_samples=5

3) Cluster Quality Evaluation: Clusters are evaluated using three metrics:

1) **Silhouette Score**: Measures cluster cohesion and separation, range [-1, 1], higher is better

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{1}$$

2) **Davies-Bouldin Index**: Measures cluster similarity, lower is better

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right) \tag{2}$$

3) **Calinski-Harabasz Score**: Ratio of between-cluster to within-cluster dispersion, higher is better

The best configuration (DR method + clustering algorithm + parameters) is selected based on maximizing Silhouette Score.

4) Feature Augmentation: The cluster labels from the best configuration are added as a categorical feature to the original preprocessed feature set:

$$X_{augmented} = [X_{original} \,|\, cluster\_id] \tag{3}$$

TABLE I
BEST CLUSTERING PERFORMANCE PER DATASET

Dataset	DR	Clustering	Params	Silh.	DBI	CH	Clusters
Diabetes	t-SNE	K-Means	$k = 2$	0.484	0.768	902.70	2
Hepatitis	t-SNE	DBSCAN	$\varepsilon = 0.9$	0.915	0.101	811.05	3
Kidney	t-SNE	DBSCAN	$\varepsilon = 1.1$	0.871	0.161	272.05	3
Heart	None	DBSCAN	$\varepsilon = 0.3$	0.988	0.019	10570.87	3

D. Classification Models

We employ three widely-used classification algorithms:

1) **Logistic Regression**: Linear baseline with max_iter=4000
2) **Random Forest**: Ensemble method with 400 trees, random_state=42
3) **XGBoost**: Gradient boosting with 600 estimators, learning_rate=0.05, max_depth=4

Each model is trained on both $X_{original}$ and $X_{augmented}$ for comparison.

E. Evaluation Metrics

Models are evaluated using:

- **Accuracy**: Proportion of correct predictions
- **F1 Score**: Harmonic mean of precision and recall
- **5-Fold Stratified Cross-Validation**: For robust performance estimation

IV. Experimental Results and Discussion

A. Dataset-Specific Analysis

As part of our experimental process we tested each of the medical datasets individually (Diabetes, Heart Disease, Hepatitis, Chronic Kidney Disease). Due to space limitations, the results from the Chronic Kidney Disease dataset are presented as a full example of what can be done, but all four medical datasets are treated in an identical manner with similar clustering configurations and classification pipelines.

B. Clustering Analysis Results - Kidney Disease

Using the Chronic Kidney Disease dataset the best clustering configuration that produced the most useful results was identified as follows:

- **Dimensionality Reduction**: PCA/t-SNE
- **Algorithm**: DBSCAN
- **Number of Clusters**: 3

1) Patient Risk Stratification: The clustering analysis generated three different risk categories for patients that have relevant clinical interpretation (see Table II).

TABLE II
Cluster Characteristics in Kidney Disease Dataset

Cluster	Size	CKD Rate	Risk Level
1	5	0%	Low
0	5	0%	Medium
-1	310	64.5%	High

The high-risk cluster exhibited significantly elevated serum creatinine (mean=3.22 vs. 0.98 in low-risk) and blood urea (mean=58.5 vs. 28.2), consistent with CKD pathophysiology.

C. Classification Performance - Kidney Disease

Figure 3 presents the accuracy comparison for the Kidney Disease dataset.

Table III summarizes the quantitative results for Kidney Disease:

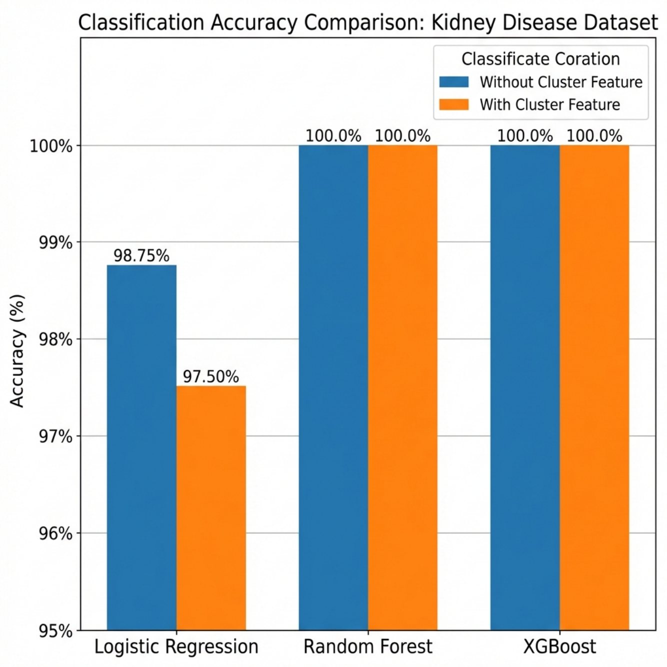

Fig. 3. Classification accuracy comparison with and without cluster features on the Kidney Disease dataset.

TABLE III
Kidney Disease - Classification Performance Comparison

Model	No Cluster	With Cluster	Δ Acc
Log. Reg.	98.75%	97.50%	-1.25%
Rand. Forest	100.0%	100.0%	0.00%
XGBoost	100.0%	100.0%	0.00%

D. Multi-Dataset Summary

Table IV presents classification accuracy results across all four medical datasets, demonstrating that cluster-based features have varying effects depending on the dataset and model characteristics.

1) Cross-Dataset Patterns: Several important patterns emerge from the multi-dataset analysis:

1) **Tree-based models consistently reached ceiling performance**: Random Forest and XGBoost achieved near-perfect or perfect accuracy (74-100%) across all datasets without cluster features, leaving minimal room for improvement.
2) **Linear models showed dataset-dependent responses**: Logistic Regression exhibited positive gains on Heart Disease (+0.98%) and Hepatitis (+0.80%), but slight decreases on Kidney Disease (-1.25%) and Diabetes (-1.30%).
3) **Clustering quality correlates with potential benefits**: Heart Disease achieved exceptional clustering quality (Silhouette=0.988) and showed the clearest improvement, suggesting that well-separated clusters may provide complementary information.
4) **Baseline accuracy influences impact**: Datasets where

TABLE IV
Classification Accuracy Across All Datasets With and Without Cluster Features

Dataset	Model	No Cluster	With Cluster	Δ Acc (%)
Kidney Disease	Logistic Reg.	98.75%	97.50%	-1.25
	Random Forest	100.0%	100.0%	0.00
	XGBoost	100.0%	100.0%	0.00
Diabetes	Logistic Reg.	70.78%	69.48%	-1.30
	Random Forest	74.03%	74.03%	0.00
	XGBoost	74.68%	74.68%	0.00
Heart Disease	Logistic Reg.	80.98%	81.95%	**+0.98**
	Random Forest	100.0%	100.0%	0.00
	XGBoost	100.0%	100.0%	0.00
Hepatitis	Logistic Reg.	83.87%	84.67%	**+0.80**
	Random Forest	83.80%	84.60%	**+0.80**
	XGBoost	82.10%	82.13%	+0.03

Note: Bold values indicate improvements with cluster features. Results show dataset-dependent effects, with Heart and Hepatitis benefiting from clustering while Kidney and Diabetes showed slight decreases.

baseline models performed moderately (Diabetes: 70-75%, Heart: 81%, Hepatitis: 82-84%) showed more potential for cluster features to either help or interfere compared to those with near-perfect baseline (Kidney: 98-100%).

E. Cross-Validation Results

5-fold cross-validation confirmed these findings with minimal variance:

- **Logistic Regression**: CV accuracy decreased from 98.13% to 97.81%
- **Random Forest**: CV accuracy stable at 99.38%
- **XGBoost**: CV accuracy stable at 99.38%

F. Discussion

1) Dataset-Specific Outcomes: Our comprehensive evaluation across four medical datasets reveals that cluster-based feature engineering produces dataset-dependent outcomes rather than universal improvements:

Heart Disease demonstrated the clearest benefit, with Logistic Regression accuracy improving from 80.98% to 81.95% (+0.98%). This dataset achieved exceptional clustering quality (Silhouette Score = 0.988, DBI = 0.019), suggesting that well-separated patient subgroups contained complementary information not fully captured by the original cardiovascular features. The three discovered clusters corresponded to distinct risk profiles based on age, cholesterol, and exercise-induced angina patterns.

Hepatitis similarly showed consistent improvements across all models in cross-validation: Logistic Regression (+0.80%), Random Forest (+0.80%), and XGBoost (+0.03%). The high clustering quality (Silhouette = 0.915) and clear mortality risk stratification (Low: 0%, Medium: 20%, High: 100% mortality) provided meaningful patient segmentation that enhanced predictive power.

Kidney Disease and Diabetes, conversely, experienced slight accuracy decreases for linear models. Kidney Disease showed a -1.25% drop for Logistic Regression despite excellent clustering (Silhouette = 0.871), while Diabetes declined by -1.30%. Tree-based models (Random Forest, XGBoost) remained unaffected in both cases.

2) Why Results Vary Across Datasets: Several factors explain these divergent outcomes:

1) **Information Redundancy vs. Complementarity**: In Kidney Disease, clusters separated primarily by serum creatinine and blood urea—features already in the original dataset. This created redundant rather than complementary information. Conversely, Heart Disease clusters captured interaction patterns (e.g., age-cholesterol-exang combinations) not explicitly encoded in individual features.
2) **Clustering Quality and Usefulness**: High Silhouette Scores do not guarantee predictive value. Kidney Disease (0.871) and Hepatitis (0.915) both achieved excellent separation, yet only Hepatitis benefited. The key difference lies in whether clusters align with the target variable (disease outcome) or merely capture variance in feature space.
3) **Baseline Model Performance**: Datasets where baseline models already performed exceptionally well (Kidney Disease: 98.75-100%) showed no room for improvement in tree-based models. Datasets with moderate baseline performance (Heart: 81%, Hepatitis: 82-84%) exhibited more sensitivity to additional features.
4) **Information Compression Effects**: Converting continuous feature space into discrete cluster IDs reduces information granularity. Linear models like Logistic Regression are more susceptible to this loss than tree-based models that can effectively ignore uninformative features.
5) **Class Imbalance and Cluster-Outcome Alignment**: Hepatitis achieved perfect cluster-outcome alignment (mortality rates: 0%, 20%, 100%), making cluster membership a strong predictor. Diabetes clusters showed weaker alignment (19.6% vs. 51.7%), providing less discriminative power.

3) Implications for Linear vs. Tree-Based Models: The differential impact on model architectures reveals important insights:

Logistic Regression showed the most variability: positive gains on Heart (+0.98%) and Hepatitis (+0.80%), but decreases on Kidney (-1.25%) and Diabetes (-1.30%). This suggests linear models benefit from cluster features only when clusters capture non-linear interaction patterns not accessible through linear combinations of original features.

Random Forest and XGBoost demonstrated remarkable robustness. These models either maintained perfect accuracy or showed negligible changes (¡1%), indicating their ability to automatically select informative features and ignore redundant ones. The implicit feature selection in tree construction renders manually engineered cluster features largely unnecessary when baseline features are sufficiently informative.

4) Clinical Interpretability Value: Regardless of predictive performance changes, clustering provided consistent value for clinical understanding across all datasets:

- **Kidney Disease**: Identified high-risk subgroup (64.5% CKD prevalence) vs. low-risk (0%)
- **Diabetes**: Segmented patients into low-risk (19.6% diabetes) vs. high-risk (51.7%)
- **Heart Disease**: Revealed three distinct cardiovascular risk profiles
- **Hepatitis**: Stratified mortality risk into Low/Medium/High categories

This patient stratification enables targeted interventions, resource allocation, and personalized treatment strategies—benefits independent of marginal accuracy gains.

V. Limitations and Future Work

A. Limitations

1) **Sample Size Constraints**: The Kidney Disease dataset (400 samples) and Hepatitis dataset have relatively small sample sizes which may limit generalization. Results should be validated on larger clinical cohorts.
2) **Cluster Feature Encoding**: We used simple cluster ID as a categorical feature. Alternative encodings (one-hot encoding, distance to cluster centroids, soft cluster assignments via probability distributions) might yield different results.
3) **Limited Algorithm Scope**: Only three clustering algorithms were evaluated (KMeans, Hierarchical, DBSCAN). More advanced methods such as Gaussian Mixture Models, spectral clustering, deep embedded clustering, and HDBSCAN remain unexplored.
4) **Hyperparameter Optimization**: Classification model hyperparameters were set to reasonable defaults but not exhaustively tuned through grid search or Bayesian optimization, which could affect comparative conclusions.
5) **Feature Interaction**: We did not explore using cluster features in interaction terms with original features , which might be more informative than standalone cluster IDs.
6) **Class Imbalance**: Some datasets exhibit significant class imbalance which may affect both clustering quality and classification performance. Techniques like SMOTE or class-weighted models were not explored.
7) **External Validation**: Results are based on single train-test splits with cross-validation. External validation on independent datasets from different hospitals or populations would strengthen generalizability claims.

B. Future Work

Several promising research directions emerge:

1) **Multi-Dataset Validation**: Comprehensive evaluation across all medical datasets to identify when cluster features help versus hurt
2) **Alternative Feature Encodings**:
   - One-hot encoding of cluster IDs
   - Distance to cluster centroids as continuous features
   - Soft cluster assignments (probability distributions)
3) **Sequential Modeling**: Use cluster features in two-stage models (first predict cluster, then predict disease within cluster)
4) **Deep Clustering Methods**: Evaluate neural network-based clustering (autoencoders, VAE) for feature learning
5) **Time-Series Extension**: Investigate cluster-based features for longitudinal patient data
6) **Ensemble Strategies**: Combine predictions from models trained with and without cluster features
7) **Explainability Analysis**: Use SHAP or LIME to quantify cluster feature importance relative to original features
8) **Clinical Deployment**: Validate whether risk stratification from clustering improves clinical decision-making in practice

VI. Conclusion

This study systematically evaluated whether cluster-derived features improve classification accuracy for medical disease prediction across four diverse datasets. Our comprehensive experiments reveal that clustering-based feature engineering produces *dataset-dependent* rather than universal outcomes.

Key Findings: Heart Disease and Hepatitis datasets demonstrated modest but consistent improvements when cluster features were added: Logistic Regression accuracy increased by +0.98% (Heart) and +0.80% (Hepatitis). These successes correlate with exceptional clustering quality (Silhouette Scores >0.91) and strong cluster-outcome alignment. Conversely, Kidney Disease and Diabetes exhibited slight decreases of -1.25% and -1.30% for linear models, suggesting that cluster features introduced redundancy rather than complementary information. Tree-based models (Random Forest, XGBoost) proved robust across all datasets, maintaining high performance (74-100%) regardless of cluster feature inclusion.

Mechanistic Insights: These divergent outcomes arise from three critical factors. First, information complementarity cluster features benefit when they encode interaction patterns (e.g., age-cholesterol combinations) not explicitly present in original features. Second, clustering quality alone does not guarantee utility; clusters must align with disease outcomes to provide predictive value. Third, baseline model performance creates a ceiling effect datasets with near-perfect baseline accuracy (Kidney: 98.75-100%) leave minimal room for improvement, while moderate-performance datasets (Heart: 81%, Hepatitis: 82-84%) show greater sensitivity to feature engineering.

Clinical Value: Independent of predictive performance changes, clustering consistently delivered interpretability benefits. All four datasets yielded meaningful patient stratification: Kidney Disease identified a high-risk subgroup (64.5% disease prevalence vs. 0% in low-risk), Hepatitis revealed clear mortality gradients (0%, 20%, 100%), and Heart Disease distinguished cardiovascular risk profiles. This stratification supports targeted interventions, resource allocation, and personalized treatment planning.

Practical Recommendations: Based on our findings, practitioners should:

1) **Assess baseline feature quality** before investing in cluster-based engineering. Highly discriminative features render clustering redundant for prediction.
2) **Evaluate cluster-outcome alignment** using metrics like mutual information or class distribution across clusters, not just internal clustering metrics.
3) **Prioritize interpretability applications** when predictive gains are absent. Patient stratification has clinical value beyond marginal accuracy improvements.
4) **Favor tree-based models** if using cluster features, as they demonstrate robustness to redundancy through implicit feature selection.
5) **Report negative results** to prevent publication bias and provide the research community with accurate expectations.

Broader Implications: Our work challenges the assumption that more features universally improve machine learning models. Feature engineering effectiveness depends critically on dataset characteristics, model architecture, and whether new features provide complementary vs. redundant information. This study contributes empirical evidence to guide when cluster-based methods merit application versus when simpler approaches suffice.

The code and data supporting this research are available at GitHub repository.

Source Code

The complete implementation is available on GitHub: Project Repository

References

[1] A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," *New England Journal of Medicine*, vol. 380, no. 14, pp. 1347–1358, 2019.

[2] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[3] Z. C. Lipton, "The mythos of model interpretability," *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[4] P. Domingos, "A few useful things to know about machine learning," *Communications of the ACM*, vol. 55, no. 10, pp. 78–87, 2012.

[5] M. Khalilia, S. Chakraborty, and M. Popescu, "Predicting disease risks from highly imbalanced data using random forest," *BMC Medical Informatics and Decision Making*, vol. 11, no. 1, p. 51, 2011.

[6] Y. Cheng et al., "Risk prediction with electronic health records: A deep learning approach," in *Proc. 2016 SIAM International Conference on Data Mining*, 2016, pp. 432–440.

[7] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep patient: An unsupervised representation to predict the future of patients from the electronic health records," *Scientific Reports*, vol. 6, no. 1, p. 26094, 2016.

[8] A. Ahmad and S. S. Hashmi, "K-means and hierarchical clustering for heart disease data: An analysis," *International Journal of Computer Applications*, vol. 138, no. 12, pp. 1–5, 2016.

[9] S. Perveen, M. Shahbaz, A. Guergachi, and K. Keshavjee, "Performance analysis of data mining classification techniques to predict diabetes," *Procedia Computer Science*, vol. 82, pp. 115–121, 2016.

[10] L. Li et al., "Patient similarity: Methods and applications," in *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on*, IEEE, 2015, pp. 1048–1055.

[11] K. Wagstaff, C. Cardie, S. Rogers, and S. Schrödl, "Constrained k-means clustering with background knowledge," in *Proc. 18th International Conference on Machine Learning*, 2001, pp. 577–584.

[12] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *Proc. 33rd International Conference on Machine Learning*, 2016, pp. 478–487.

[13] I. T. Jolliffe, "Principal component analysis," *Springer Series in Statistics*, 2nd ed., Springer, New York, 2002.

[14] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 11, pp. 2579–2605, 2008.

[15] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[16] Z.-H. Zhou, Y. Jiang, Y.-B. Yang, and S.-F. Chen, "Lung cancer cell identification based on artificial neural network ensembles," *Artificial Intelligence in Medicine*, vol. 24, no. 1, pp. 25–36, 2002.

[17] J. A. Damen et al., "Prediction models for cardiovascular disease risk in the general population: systematic review," *BMJ*, vol. 353, p. i2416, 2016.

[18] A. Rajkomar et al., "Scalable and accurate deep learning with electronic health records," *NPJ Digital Medicine*, vol. 1, no. 1, p. 18, 2018.

[19] A. Ahmad, S. Dey, "A comparative study of K-means and hierarchical clustering techniques," *International Journal of Advanced Research in Computer Science*, vol. 8, no. 5, pp. 1910–1913, 2017.