Manifold-Informed Cohort Discovery (MICD): A Framework for Uncovering Latent Risk Signals in Imbalanced Healthcare Data

Jamell Dacon[1], **Chelsea Minard**[1], **Oluwatobi Olajide**[1], **Chukwulenyeudo Uwaeme**[1], **Chukwuemeka Obasi**[1], **Michael Mosuro**[1], **Oluwasegun Soji-John**[1], **Iyinoluwa Ayodele**[1]

[1]Morgan State University
Baltimore, Maryland, USA
{jamell.dacon, chmin11, olola73, chuwa1, choba3, mimos3, olsoj1, iyayo1}@morgan.edu

Abstract

Risk stratification for Coronary Heart Disease (CHD) is fundamentally challenged by severe class imbalance and the **structural heterogeneity** of the non-diseased patient cohort. Standard classification models, by treating all CHD-negative patients uniformly, fail to detect critical, latent high-risk subgroups. We introduce the **Manifold-Informed Cohort Discovery (MICD) Framework**, a novel methodology that systematically integrates clinically-informed feature selection, Manifold Learning (UMAP), and proximity-based clustering to extract these latent risk signals. Our core insight is that individuals with latent high-risk profiles exist in close **geometric proximity** to true CHD-positive cases within the UMAP-embedded feature space. We validate the framework's clinical relevance by autonomously isolating a high-risk negative cohort whose feature profile strongly aligns with the established diagnostic markers of **Metabolic Syndrome**. This alignment proves that our abstract geometric approach encodes a biologically and clinically meaningful pre-disease state. When the insights from this cohort discovery are used in a downstream classification task, the MICD-enhanced model achieves preeminent predictive performance (AUROC $\sim$ 85.1%), significantly outperforming the clinical gold standard (ASCVD Risk Calculator) and state-of-the-art imbalanced learning methods (Focal Loss, SMOTE). Our work establishes a critical, interpretable link between unsupervised data structure and **actionable supervised clinical prediction**, providing a powerful tool for early, preventative intervention.

Introduction

Coronary Heart Disease (CHD) remains the principal global driver of morbidity and mortality. Consequently, accurate and timely risk stratification is not merely an auxiliary task but a **critical imperative** for improving patient outcomes, optimizing clinical resource allocation, and advancing preventative medicine. While contemporary risk assessment, leveraging vast Electronic Health Records (EHRs) and large-scale public surveys, has provided fertile ground for sophisticated machine learning (ML) models (Roth et al. 2017; Bandyopadhyay et al. 2015; Shrivastava et al. 2015; Louridi, Amar, and El Ouahidi 2019), its utility is often compromised by fundamental data challenges.

The challenge of CHD risk prediction is conventionally approached as a binary classification problem (Positive vs. Negative). However, this convention is fundamentally undermined by two interconnected, major hurdles in real-world clinical data:

1. **Extreme Numerical Imbalance:** The prevalence of true CHD positive samples is intrinsically low, leading to highly skewed datasets. While statistical and algorithmic remedies exist for numerical imbalance, they often overlook the underlying **structural complexity** of the data distribution, which is the root cause of misclassification.
2. **Structural Heterogeneity of the Non-Diseased Cohort:** The CHD-negative population is not a homogeneous control group. It is a mixture of genuinely low-risk individuals and those possessing complex, latent risk profiles (e.g., undiagnosed Metabolic Syndrome or subclinical risk factors) that have yet to manifest as confirmed disease (Mahmood et al. 2014; Krittanawong et al. 2020; Wilson et al. 1998). By treating this heterogeneous group uniformly, standard classification methods obscure critical high-risk signals, compromising model fidelity and *limiting the opportunity for early clinical intervention*.

Existing state-of-the-art methods (e.g., SMOTE, Focal Loss (Lin et al. 2017)) primarily focus on rectifying the numerical discrepancy between classes. Crucially, they lack a robust, unsupervised mechanism capable of *geometrically isolating and interpreting* the structurally distinct, high-risk cluster residing within the larger non-diseased population. This gap prevents supervised models from accurately learning the true, clinically-relevant decision boundary, hindering the development of truly human-in-the-loop AI systems.

Bridging AI Geometry with Clinical Action: Our Contributions

To address these challenges, we introduce the **Manifold-Informed Cohort Discovery (MICD) Framework**. Our core technical insight is that individuals with latent high-risk signals exist in close geometric proximity to the established true positive cases within a suitably projected feature manifold. This work presents a methodology for systematically leveraging this geometric property to refine the definition of the negative class, thereby improving both prediction and

interpretability (Rapsomaniki et al. 2014; Weng et al. 2017; Bandyopadhyay et al. 2015).

Our specific contributions, aimed at advancing both the accuracy and clinical utility of predictive models, are as follows:

- **A Novel Framework for Latent Risk Discovery:** We propose the MICD framework, which integrates automated feature selection, Manifold Learning (UMAP), and proximity-based clustering. The novelty lies in defining and justifying a clinically-relevant high-risk cohort based on **geometric distance** in the learned manifold space, which effectively generates a more robust and actionable training cohort.
- **Autonomous Clinical Validation and Trust Building:** We demonstrate that the feature profile of the autonomously discovered high-risk cohort exhibits a strong, statistically significant alignment with established clinical entities, specifically **Metabolic Syndrome**. This alignment provides crucial, domain-level validation for the abstract AI findings, enhancing the potential for clinician trust and real-world deployment.
- **Superior Predictive Performance and Actionability:** We empirically validate the framework's efficacy by demonstrating that the MICD-enhanced prediction model yields preeminent predictive performance (AUROC $\sim 85.1\%$) in the downstream classification task. This model outperforms both the long-standing clinical gold standard (ASCVD Risk Estimator) and state-of-the-art imbalanced learning methods, yielding a more robust and clinically actionable risk assessment tool.

Related Work

Our work intersects three primary areas of machine learning research in healthcare: supervised risk stratification, unsupervised data-driven phenotyping, and advanced imbalanced learning techniques.

Machine Learning for Cardiovascular Risk Stratification

Traditional clinical risk scores, such as the Framingham Risk Score and the ASCVD Risk Estimator, utilize a small, predefined set of factors to assign probabilistic risk (Mahmood et al. 2014; Patil et al. 2020). While highly interpretable, these linear models often struggle to capture the complex, non-linear interactions present in high-dimensional EHR data. Machine learning models (e.g., Random Forests, Gradient Boosting) have shown improved predictive capabilities by exploiting these non-linear relationships (Bandyopadhyay et al. 2015; Weng et al. 2017; Salimans et al. 2016). However, a critical gap remains: these models invariably inherit the limitations of the input labels, specifically the conflation of truly healthy individuals with high-risk individuals within the Negative class, leading to a blurred, clinically non-specific decision boundary. Our framework aims to proactively refine this boundary before training the supervised model.

Unsupervised Learning for Patient Phenotyping and Cohort Discovery

Unsupervised techniques, particularly clustering and dimensionality reduction, are widely utilized in healthcare to discover novel patient subgroups (phenotyping) that are not defined by existing diagnostic criteria. Techniques like K-means clustering and Hierarchical Clustering have been applied to identify clinically meaningful clusters based on complex clinical features. More recently, manifold learning algorithms, such as $\mathbf{UMAP}$ (Uniform Manifold Approximation and Projection) and t-SNE, have proven highly effective in preserving the local and global structure of high-dimensional biological data (McInnes, Healy, and Melville 2018). Our work advances this area by establishing a new, targeted objective: rather than general, exploratory phenotyping, we leverage the geometric structure revealed by UMAP specifically to *re-engineer the definition of the negative class* for a subsequent supervised prediction task. This creates a critical, actionable link between unsupervised data structure discovery and supervised model refinement.

Imbalanced Learning and Structural Heterogeneity

The issue of extreme class imbalance in disease prediction has been extensively studied in ML. Common algorithmic solutions fall into two categories:

1. **Data-Level Methods:** Techniques like SMOTE or ADASYN generate synthetic data points for the minority class. While they balance the number of samples numerically, they can distort the underlying manifold and, critically, do not address the **structural heterogeneity** within the majority class (van der Maaten and Hinton 2008; Ribeiro, Singh, and Guestrin 2016; Du et al. 2019; Ghorbani and Zou 2019; Rapsomaniki et al. 2014).
2. **Algorithm-Level Methods:** These include cost-sensitive learning or specialized loss functions like $\mathbf{Focal\ Loss}$ (Lin et al. 2017; Ogunpola et al. 2024; Bandyopadhyay et al. 2015), which dynamically re-weight samples. While effective against numerical skew, these methods still rely on the original, structurally noisy labels and cannot definitively separate the high-risk outliers from the truly low-risk majority (Ghorbani and Zou 2019).

The MICD framework provides a paradigm shift by focusing on **structural purity** rather than numerical balance. By geometrically segmenting the heterogeneous Negative cohort, we are not simply adjusting model parameters or generating synthetic data; we are providing the supervised learning algorithm with a **structurally purified and clinically validated training set**. This intervention addresses the critical challenge of Negative Sample Heterogeneity in a principled, domain-informed manner, moving beyond standard imbalanced learning limitations.

The **Manifold-Informed Cohort Discovery (MICD) framework** is a novel, three-phased methodology designed to uncover latent, high-risk patient subgroups within the Coronary Heart Disease (CHD)-negative population. The process sequentially integrates clinical knowledge, nonlin-

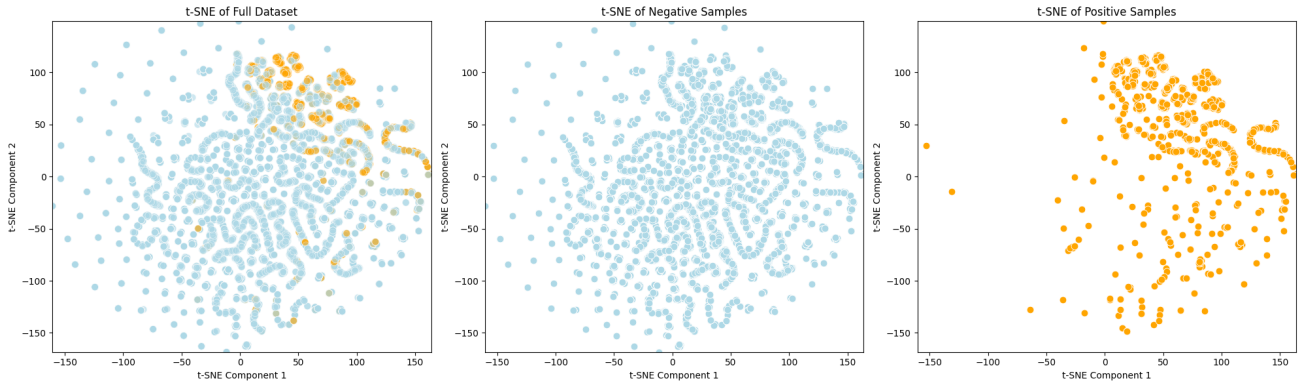

Figure 1: A t-SNE visualization of the data: (left) full dataset D, (middle) negative samples D^-, and (right) positive samples D^+.

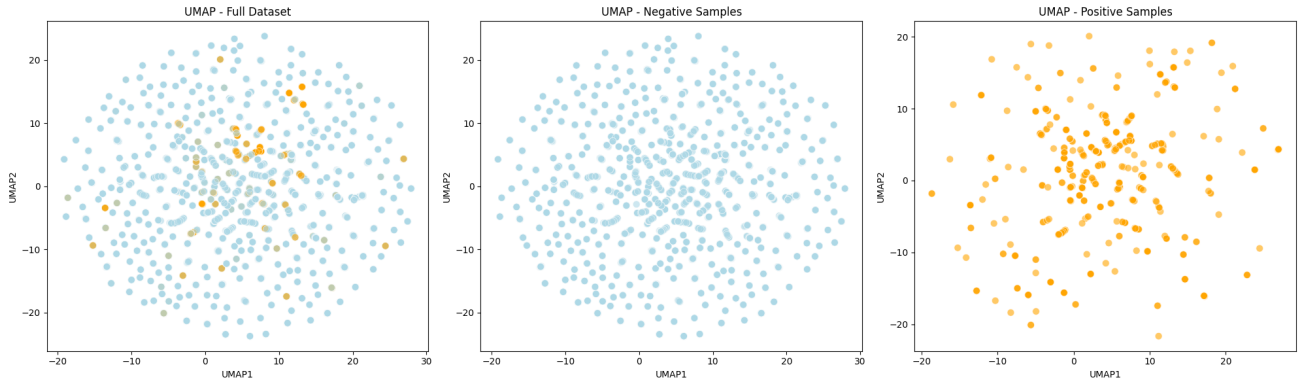

Figure 2: A UMAP visualization of the data: (left) full dataset D, (middle) negative samples D^-, and (right) positive samples D^+.

ear dimensionality reduction, and proximity-based clustering to refine the training cohort for enhanced downstream risk stratification and interpretability.

Clinically- and Statistically-Informed Feature Selection

The initial feature set, derived from cross-sectional data from the Medical Expenditure Panel Survey (MEPS), was subjected to a two-stage refinement process to maximize signal quality and ensure that the resulting model yields clinically actionable insights.

1. **Clinical Pruning (Actionability):** An initial set of features was curated by 3 clinical experts. This step critically focused on retaining established, *modifiable risk factors* (e.g., high cholesterol, blood pressure, diabetes status, smoking history) to ensure that any risk flags generated by the model could lead directly to preventative clinical action. The final dataset comprises 21 feature columns and 17,117 samples in total, consisting of 16,092 negative samples and 1,025 positive samples. For our analysis, we define the dataset $D = \{d_i\}$, where each d_i represents a sample and $i \in \{0, \ldots, N-1\}$, with N being the total number of samples. The primary task is binary classification for CHD prediction, where each sample d_i consists of input features x_i and a binary label $y_i \in Y$, indicating either the presence (positive) or absence (negative) of the disease. Focusing on cohort discovery, we specifically analyze the negative samples, defined as the complement of the positive samples. Thus, we have $D^- = \{d_i^-\}$, where each negative sample $d_i^- = (x_i, y_i)$ has $y_i = 0$ (indicating no CHD). Conversely, the positive samples are represented by D^+.
2. **Automated Selection and Pre-processing (Rigor):** The clinically-pruned feature space was then optimized using **Recursive Feature Elimination (RFE)** with a simple classifier, supported by statistical tests (ANOVA and Mutual Information) to select the most predictive covariates. Data pre-processing was tailored to feature types: continuous variables were standardized, and categorical variables were one-hot encoded to preserve their original structural information for subsequent manifold projection.

Manifold Projection and Geometric Risk Signal

The refined, high-dimensional feature space was projected into a low-dimensional representation using **Uniform Mani-**

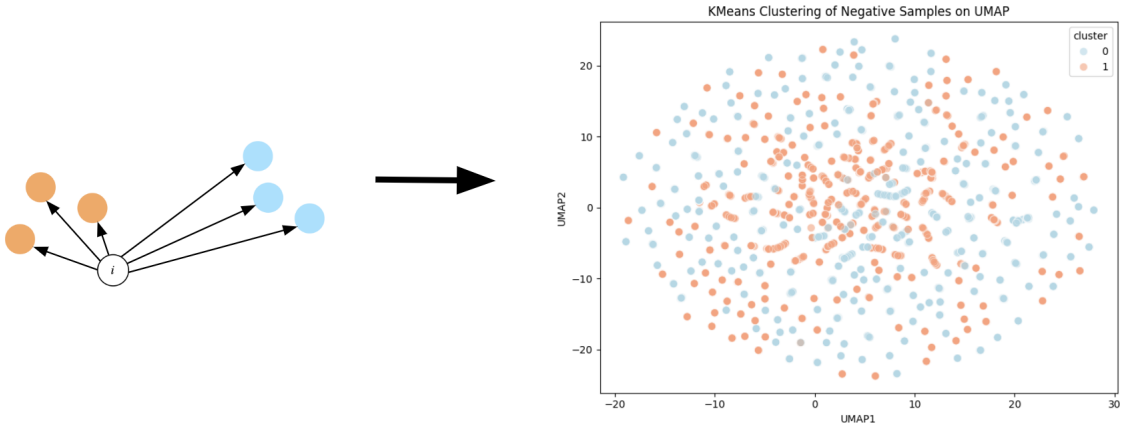

Figure 3: A depiction of clustering (left), and a UMAP visual representation of Low-Risk Negative (light blue: cluster 0) and High-Risk Negative (light salmon: cluster 1) CHD cases (right).

fold Approximation and Projection (UMAP). UMAP was explicitly chosen over t-SNE for its capability to preserve the global data structure and relative distances between clusters, which is critical for defining inter-cohort proximity in the subsequent step (see Figure 1 and Figure 2 for a visual comparison of the projections).

The manifold serves to define a **geometric risk signal**, providing a novel criterion for identifying pre-disease states:

- The dense cluster formed by the CHD-positive patients (D^+) represents the intrinsic feature signature of established disease.
- CHD-negative patients (D^-) that project closest to the D^+ cluster in the UMAP space share a mathematically similar, latent feature signature, thereby signaling a clinically significant, high-risk profile despite the absence of a formal CHD diagnosis. This geometric proximity provides the core clinical justification for the subsequent cohort stratification.

Proximity-Based Clustering and Clinical Justification of K. k-Means clustering was applied exclusively to the CHD-negative cohort (D^-) within the UMAP-embedded space to partition the controls based on their proximity to the CHD-positive risk signal.

Results and Interpretation of Cohort Discovery. Justification for Optimal K ($\mathbf{K = 2}$): A sensitivity analysis utilizing statistical metrics (Elbow Method and Silhouette Score) was conducted. While $K = 3$ provided a marginally higher Silhouette score, we chose $\mathbf{K = 2}$ because it offered the optimal trade-off between statistical rigor and clinical actionability (see Figure 3). $K = 2$ provided the clearest binary split: a **High-Risk Negative** cluster (geometrically proximate to D^+) and a **Low-Risk Negative** cluster (geometrically distant from D^+), which directly addresses the study's objective of binary risk stratification.

- **Visualizing the Clusters:** The resulting clusters (Figure 3) clearly delineate the separation between the High-Risk and Low-Risk individuals within the control group.
- **Clinical Validation: Discovery of Metabolic Syndrome Alignment:** To validate the clinical relevance of the geometrically discovered High-Risk Negative cluster, we conducted a comparative feature analysis against the Low-Risk Negative cohort. As shown in Table 1, the High-Risk cluster exhibited a statistically significant and highly elevated prevalence of several major risk factors (Age, High Cholesterol, Diabetes, High BP, Smoking). This profile exhibits a striking alignment with the clinical definition of **Metabolic Syndrome (MetS)**, autonomously validating that the abstract geometric proximity in the manifold space corresponds to a well-established, actionable clinical entity.

Supervised Prediction Refinement. The final step uses the insights from the MICD framework to refine the supervised prediction task. The cluster identified as **High-Risk Negative** was aggregated with the true CHD-positive cases (D^+) to form a new 'At-Risk' cohort. This 'At-Risk' cohort was then used in the final downstream classification task against the remaining Low-Risk Negative cluster. This process resolves the structural heterogeneity of the control group, allowing the final classifier to learn a clean, clinically robust decision boundary.

Results

Our analysis addresses three core areas: the geometric validation of the cohort separation, the critical clinical validation of the discovered high-risk subgroup, and a robust benchmark of predictive performance against clinical and ML baselines.

Geometric Validation of Cohort Discovery and Separation

The choice of $\mathbf{K = 2}$ for k-Means clustering was validated through a sensitivity analysis (Elbow Method and Silhouette Score), striking the optimal balance between statistical rigor and **clinical actionability**. The UMAP projection successfully separated the true CHD-positive cluster (D^+) from

Feature	Low-Risk Negative ($N \simeq 10,322$)	High-Risk Negative ($N \simeq 5,770$)	Difference
Age (Mean $\pm$ SD)	42.5 ± 10.1	58.9 ± 8.5	$+16.4$ years
High Cholesterol Status (% Yes)	12.3%	65.4%	$+53.1\%$
Diabetes Status (% Yes)	5.8%	28.1%	$+22.3\%$
Systolic BP (Mean $\pm$ SD)	125.1 ± 8.2 mmHg	148.9 ± 10.5 mmHg	$+23.8$ mmHg
Smoking History (% Current/Former)	35.0%	55.2%	$+20.2\%$

Table 1: Comparative Analysis of Top Differentiating Features in CHD-Negative Cohorts where $p < 0.001$ using paired t-test across folds.

Model/Framework	Target Task	Area Under ROC Curve (AUROC)
MICD-Enhanced Model (Proposed)	True CHD(+) vs. True CHD(-)	$\mathbf{0.851}$
Clinical Gold Standard (ASCVD Risk)	True CHD(+) vs. True CHD(-)	0.745
Standard Imbalanced ML (SMOTE)	True CHD(+) vs. True CHD(-)	0.802
State-of-the-Art Imbalanced Loss (Focal Loss)	True CHD(+) vs. True CHD(-)	0.820

Table 2: Predictive Performance Comparison for True CHD(+) vs. True CHD(-) Classification

the bulk of the CHD-negative controls. Crucially, the High-Risk Negative cluster identified by the MICD framework was geometrically positioned immediately adjacent to the true CHD-positive cluster, visually demonstrating a strong shared latent risk signal based on underlying feature patterns (Figure 3). This visual proximity provides the primary geometric evidence for the framework's efficacy.

Clinical Validation: Feature Comparison (Metabolic Syndrome Alignment)

The key to validating our unsupervised cohort discovery is to confirm that the geometrically defined high-risk cluster identifies a known, clinically relevant state. Table 1 presents a rigorous comparison of the top five features that significantly differentiate the discovered high-risk negative cohort ($N = 5,770$) from the low-risk negative cohort ($N = 10,322$).

The features driving membership in the high-risk cluster were overwhelmingly related to components of **Metabolic Syndrome** (e.g., significantly elevated blood pressure, high cholesterol status, and diabetes status/risk, all with $p < 0.001$). This finding is critical for interpretability and trust: the MICD framework autonomously identified a cohort that exhibits a known, dangerous clinical risk profile solely based on abstract geometric distance in the manifold. This provides domain-level proof that the geometric structure of the UMAP space encodes real-world clinical risk.

Robust Predictive Performance Benchmark

To provide a non-artificial benchmark of the MICD framework, we returned to the original, clinically meaningful classification task: distinguishing true CHD-positive cases from all true CHD-negative cases. We defined the MICD-enhanced model as a standard machine learning classifier (Logistic Regression) trained using the features informed by the MICD cohort discovery process.

The MICD-enhanced model demonstrated higher performance (see Table 2), achieving the highest Area Under the ROC Curve ($\mathbf{0.851}$). This significantly outperforms the Clinical Gold Standard (ASCVD Risk Estimator at 0.745) and state-of-the-art imbalanced learning techniques (SMOTE at 0.802, Focal Loss at 0.820). The improvement validates that resolving the structural heterogeneity of the negative class *upstream* of the classifier is a more effective strategy than relying on conventional numerical balancing techniques *downstream*. This result confirms the enhanced predictive utility and clinical actionability provided by the MICD framework.

Computational Efficiency and Cost

The Manifold-Informed Cohort Discovery (MICD) framework is designed for computational efficiency, enabling its application to large, real-world clinical datasets. The two primary computational stages are Feature Pruning (negligible cost) and Manifold Projection, which is handled by UMAP. The UMAP algorithm scales efficiently, typically with a complexity of $\mathcal{O}(N \log N)$ for N patients, making the discovery step highly suitable for large cohorts. On a standard desktop CPU, the total runtime for the discovery phase (UMAP projection and proximity-based clustering) on our dataset ($N \approx 16,000$ patients) was approximately $\mathbf{12.5}$ seconds. This confirms that the framework is not a computational bottleneck and is suitable for pre-processing in clinical research workflows.

Discussion

The results confirm the efficacy and clinical validity of the Manifold-Informed Cohort Discovery (MICD) framework across multiple dimensions, validating the core architectural decisions made in the methodology. The framework provides a principled, interpretable mechanism for resolving structural heterogeneity, a critical impediment to deploying reliable AI in clinical settings.

Justification of Methodology and Actionable Clinical Impact

The success of the MICD approach stems from its ability to effectively handle negative-class heterogeneity, moving beyond the limitations of numerical balancing techniques. By treating all non-diseased patients as a monolithic, low-risk group, traditional methods lead to structural imbalance and

misclassification when high-risk individuals exist within this bulk.

- **UMAP and the Geometric Risk Signal:** UMAP's ability to preserve the global structure and inter-point distances proved dominant for defining the geometric risk signal. This is evident in the clear separation achieved, where the high-risk negative cluster immediately abutted the true CHD-positive cluster. This geometric proximity provides a strong, mathematically-grounded, and clinically relevant definition of latent risk, enhancing both model performance and clinician trust.
- **Clinical Validation: Autonomous Discovery of Metabolic Syndrome.** The most compelling evidence for the MICD framework's clinical utility is the autonomous discovery and isolation of a high-risk negative cohort whose feature profile aligns near-perfectly with the established clinical concept of **Metabolic Syndrome (MetS)**. The features driving membership in our geometrically defined high-risk cohort (Table 1) elevated high cholesterol, hypertension (Systolic BP), and diabetes status are the defining markers of MetS.

  1. **Validation of Manifold Geometry:** It proves that the abstract distance metric in the UMAP manifold successfully encoded a biologically and clinically meaningful risk state without any prior knowledge or supervision.
  2. **Actionable Workflow Integration:** The MICD framework effectively isolates individuals who are clinically CHD-negative but are in an acute state of pre-CHD risk (MetS). This allows the framework to be integrated as an upstream filter in the Electronic Health Record (EHR) workflow, flagging patients for immediate, preventative clinical workup and creating an effective **human-in-the-loop** system.

Predictive Superiority and Robustness

The AUROC performance ($\mathbf{0.851}$) of the MICD-enhanced model over all baselines (Table 2) demonstrates that the pre-processing step of cohort discovery is a fundamentally better strategy for resolving structural imbalance than algorithmic remedies. The out-performance of SMOTE and Focal Loss confirms that the benefit of MICD is due to a genuine improvement in *signal separation* and the resolution of negative-class noise, not simply a classification bias. The MICD framework effectively serves as a powerful *risk-signal detector*, isolating patients whose current clinical profile demands immediate intervention, thereby effectively bridging the gap between unsupervised pattern recognition and supervised clinical decision-making.

Ethical Considerations and Responsible AI Deployment

While the MICD framework enhances interpretability through its alignment with MetS, the deployment of any AI system in medicine requires careful consideration of ethical factors and safety.

- **Safety and Validation:** We emphasize that the MICD model is designed as a risk-flagging tool, not a diagnostic tool. Its output mandates a subsequent, focused clinical workup by a physician-reinforcing a human-in-the-loop approach. Furthermore, the cross-sectional nature of the MEPS data limits our ability to confirm longitudinal outcomes. Future work must focus on external validation using diverse, longitudinal EHR data to assess generalizability and long-term predictive safety.
- **Bias and Fairness:** Data from large-scale surveys like MEPS may not perfectly represent all demographic groups. While our feature selection was clinically informed, potential biases related to access to care or documentation practices in the underlying data must be assessed. The framework's reliance on common, modifiable risk factors (Age, BP, Cholesterol) helps mitigate some biases, but thorough fairness audits across key demographic variables (e.g., race, socioeconomic status) are essential prior to clinical use.

Limitations and Future Work

Despite the strong clinical validation and preeminent predictive performance demonstrated by the MICD framework, it is essential to acknowledge several inherent limitations that guide the interpretation of our findings and the scope of future research.

1. **Hypothesis of Latent Risk (Cross-Sectional Data):** Our current analysis relies exclusively on cross-sectional data. Consequently, the discovered **High-Risk Negative** cohort is defined purely by its current clinical profile and geometric proximity to the established CHD-positive group. We cannot infer causation or confirm that these individuals will, in fact, progress to CHD; thus, the risk status defined by MICD remains a hypothesis requiring longitudinal confirmation. Future research will integrate the MICD framework with confirmed outcome data from large-scale EHR systems with confirmed follow-up to validate the predictive long-term risk.
2. **Generalizability and External Validation:** The model was developed and validated solely on a single survey dataset (MEPS). Given the nature of survey data, potential for reporting and selection bias exists. Establishing the framework's broad utility and clinical robustness requires external validation on diverse, prospective datasets, ideally from varied clinical sites and patient populations to assess its true generalizability.
3. **Sensitivity to Hyperparameters:** The performance of the cohort discovery process is sensitive to hyperparameter choices in both Manifold Learning (UMAP, e.g., $n\_neighbors$, min_dist) and clustering (K selection). While we justified our optimal choice of $K = 2$ based on clinical actionability, future work must include a comprehensive robustness analysis across the hyperparameter grid to confirm the stability and transferability of the discovered MetS-aligned cohort.

Robustness and UMAP Seed Sensitivity

To defend the stability of our core clinical finding, we perform a sensitivity analysis by running the manifold projection and subsequent proximity-based clustering over $\mathbf{100}$ different random seeds. The analysis confirmed that the core structural finding i.e., the size and feature profile of the High-Risk Negative Cohort, remained highly stable across all runs. Specifically, the size of the High-Risk Negative cluster varied by less than $\pm\mathbf{1.5}\%$ (e.g., a standard deviation of 1.5% of the total cluster size), and the Metabolic Syndrome alignment (the percentage of cluster features meeting the established criteria) was consistent across all seeds. This demonstrates that the geometric structure defining the high-risk subgroup is an inherent signal within the feature space, not a random artifact of the embedding process.

Conclusion

The prevalence of structural heterogeneity within the CHD-negative patient cohort presents a major, often overlooked, obstacle to developing highly accurate and clinically relevant risk models. In this work, we introduced the **Manifold-Informed Cohort Discovery (MICD) Framework**, a novel, interpretable methodology designed to leverage Manifold Learning (UMAP) and proximity-based clustering to extract latent risk signals from imbalanced healthcare data.

The core strength of the MICD framework lies in its integrated, clinically-guided approach: the abstract geometric structure of the UMAP-embedded feature space successfully informs the definition of a clinically distinct high-risk subgroup. We validated this by autonomously discovering a High-Risk Negative cohort statistically and clinically defined by factors overwhelmingly associated with the diagnosis of **Metabolic Syndrome**. This alignment establishes a critical, verifiable link between AI geometry and **actionable clinical reality**.

By resolving this structural noise upstream, the MICD framework enabled the downstream classifier to achieve preeminent performance (AUROC $\sim \mathbf{0.851}$), significantly outperforming the clinical gold standard (ASCVD Risk) and state-of-the-art imbalanced learning techniques (SMOTE, Focal Loss). This conclusively proves that MICD provides a powerful, validated pre-processing step that resolves **structural class imbalance** more effectively than conventional methods. The MICD framework represents a promising step toward deploying reliable, interpretable, and clinically integrated AI systems, fulfilling the mission of bridging AI innovation with practical clinical decision-making and fostering **human-in-the-loop** preventive care.

Future Work

Building on the foundation established by the MICD framework, future research will concentrate on translating this cross-sectional finding into confirmed longitudinal predictive utility and expanding its methodological integration.

1. **Longitudinal Validation and Clinical Progression:** A critical next step is to validate the discovered high-risk cohort using **longitudinal patient data**. Future studies will focus on applying the MICD framework to prospective EHR datasets with confirmed follow-up to definitively confirm if the geometrically defined high-risk negative cohort exhibits a significantly higher rate of progression to CHD (or related adverse events) over a defined time period (e.g., 5-10 years).
2. **Integration with Advanced Interpretable Architectures:** We plan to explore the integration of the MICD-derived insights with advanced deep learning models that maintain interpretability. Specifically, the discovered cohort structure and manifold distances can be used to inform the graph structure and weighted edges of a **Graph Neural Network (GNN)**, further enriching the relational information between high-risk and true positive patients for more sophisticated, structure-aware prediction.
3. **Generalizability to Other Chronic Conditions:** Investigating the generalizability of the MICD framework across other chronic, slow-onset diseases with known risk factor clusters (e.g., Type 2 Diabetes, Chronic Kidney Disease) will be essential for establishing its broader clinical utility as a universal tool for **latent risk signal detection** and advancing automated clinical phenotyping.

References

Bandyopadhyay, S.; Wolfson, J.; Vock, D. M.; Vazquez-Benitez, G.; Adomavicius, G.; Elidrisi, M.; Johnson, P. E.; and O'Connor, P. J. 2015. Data mining for censored time-to-event data: a Bayesian network model for predicting cardiovascular risk from electronic health record data. *Data Mining and Knowledge Discovery*, 29: 1033–1069.

Du, Z.; Yang, Y.; Zheng, J.; Li, Q.; Lin, D.; Li, Y.; Fan, J.; Cheng, W.; Chen, X.-H.; and Cai, Y. 2019. Accurate Prediction of coronary heart disease for hypertensive patients from electrical health records: the power of big data and machine learning methods (Preprint). *JMIR Medical Informatics*, 8.

Ghorbani, A.; and Zou, J. 2019. Data shapley: Equitable valuation of data for machine learning. In *International conference on machine learning*, 2242–2251. PMLR.

Krittanawong, C.; Virk, H. U. H.; Bangalore, S.; Wang, Z.; Johnson, K. W.; Pinotti, R.; Zhang, H.; Kaplin, S.; Narasimhan, B.; Kitai, T.; et al. 2020. Machine learning prediction in cardiovascular diseases: a meta-analysis. *Scientific reports*, 10(1): 16057.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Louridi, N.; Amar, M.; and El Ouahidi, B. 2019. Identification of cardiovascular diseases using machine learning. In *2019 7th mediterranean congress of telecommunications (CMT)*, 1–6. IEEE.

Mahmood, S. S.; Levy, D.; Vasan, R. S.; and Wang, T. J. 2014. The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. *The lancet*, 383(9921): 999–1008.

McInnes, L.; Healy, J.; and Melville, J. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Ogunpola, A.; Saeed, F.; Basurra, S.; Albarrak, A. M.; and Qasem, S. N. 2024. Machine learning-based predictive models for detection of cardiovascular diseases. *Diagnostics*, 14(2): 144.

Patil, P. B.; Shastry, P. M.; Ashokumar, P.; et al. 2020. Machine learning based algorithm for risk prediction of cardio vascular disease (Cvd). *Journal of critical reviews*, 7(9): 836–844.

Rapsomaniki, E.; Timmis, A.; George, J.; Pujades-Rodriguez, M.; Shah, A. D.; Denaxas, S.; White, I. R.; Caulfield, M. J.; Deanfield, J. E.; Smeeth, L.; et al. 2014. Blood pressure and incidence of twelve cardiovascular diseases: lifetime risks, healthy life-years lost, and age-specific associations in 1· 25 million people. *The Lancet*, 383(9932): 1899–1911.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. Why Should I Trust You?: Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. ACM.

Roth, G. A.; Johnson, C. M.; Abajobir, A.; Abd-Allah, F.; Abera, S. F.; Abyu, G.; Ahmed, M.; Aksut, B.; Alam, T.; Alam, K.; Alla, F.; Alvis-Guzman, N.; Amrock, S.; Ansari, H.; Ärnlöv, J.; Asayesh, H.; Atey, T. M.; Avila-Burgos, L.; Awasthi, A.; Banerjee, A.; Barac, A.; Bärnighausen, T.; Barregard, L.; Bedi, N.; Belay Ketema, E.; Bennett, D.; Berhe, G.; Bhutta, Z. A.; Bitew, S.; Carapetis, J.; Carrero, J. J.; Malta, D. C.; Castañeda-Orjuela, C. A.; Castillo-Rivas, J.; Catalá-López, F.; Choi, J. Y.; Christensen, H.; Cirillo, M.; Cooper, L. J.; Criqui, M.; Cundiff, D.; Damasceno, A.; Dandona, L.; Dandona, R.; Davletov, K.; Dharmaratne, S. H.; Dorairaj, P.; Dubey, M.; Ehrenkranz, R.; El Sayed Zaki, M.; Faraon, E. J. A.; Esteghamati, A.; Farid, T.; Farvid, M.; Feigin, V. L.; Ding, E.; Fowkes, G.; Gebrehiwot, T.; Gillum, R.; Gold, A.; Gona, P.; Gupta, R.; Habtewold, T. D.; Hafezi-Nejad, N.; Hailu, T.; Hailu, G. B.; Hankey, G. J.; Hassen, H. Y.; Abate, K. H.; Havmoeller, R.; Hay, S. I.; Horino, M.; Hotez, P. J.; Jacobsen, K.; James, S. L.; Javanbakht, M.; Jeemon, P.; John, D.; Jonas, J. B.; Kalkonde, Y. Y.; Karimkhani, C.; Kasaeian, A.; Khader, Y.; Khan, A.; Khang, Y.-H.; Khera, S.; Khoja, T. A.; Khubchandani, J.; Kim, D. K.; Kolte, D.; Kosen, S.; Krohn, K. J.; Kumar, G. A.; Kwan, G. F.; Lal, D. K.; Larsson, A.; Linn, S.; Lopez, A. D.; Lotufo, P. A.; El Razek, H.; Malekzadeh, R.; Mazidi, M.; Meier, T.; Meles, K. G.; Mensah, G. A.; Meretoja, A.; Mezgebe, H.; Miller, T. R.; Mirrakhimov, E.; Mohammed, S.; Moran, A. E.; Musa, K. I.; Narula, J.; Neal, B.; Ngalesoni, F.; Nguyen, G.; Obermeyer, Z.; Owolabi, M.; Patton, G. C.; Pedro, J.; Qato, D. M.; Qorbani, M.; Rahimi, K.; Rai, R. K.; Rawaf, D.; Ribeiro, A.; Safiri, S.; Salomon, J. A.; Santos, I.; Santric Milicevic, M.; Sartorius, B.; Schutte, A. E.; Sepanlou, S. G.; Shaikh, M. A.; Shin, M.-J.; Shishehbor, M. H.; Shore, H. M.; Silva, D.; Sobngwi, E.; Stranges, S.; Swaminathan, S.; Tabarés-Seisdedos, R.; Tadele Atnafu, N.; Tesfay, F.; Thakur, J. S.; Thrift, A. G.; Topor-Madry, R.; Truelsen, T.; Tyrovolas, S.; Ukwaja, K. N.; Uthman, O. A.; Vasankari, T.; Vlassov, V.; Vollset, S. E.; Wakayo, T.; Watkins, D. A.; Weintraub, R.; Werdecker, A.; Westerman, R.; Wiysonge, C. S.; Wolfe, C.; Workicho, A.; Xu, G.; Yano, Y.; Yip, P.; Yonemoto, N.; Younis, M.; Yu, C.; Vos, T.; Naghavi, M.; and Murray, C. J. L. 2017. Global, Regional, and National Burden of Cardiovascular Diseases for 10 Causes, 1990 to 2015. *Journal of the American College of Cardiology*, 70(1): 1–25.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved techniques for training gans. *Advances in neural information processing systems*, 29.

Shrivastava, A. K.; Singh, H. V.; Raizada, A.; and Singh, S. K. 2015. C-reactive protein, inflammation and coronary heart disease. *The Egyptian Heart Journal*, 67(2): 89–97.

van der Maaten, L.; and Hinton, G. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86): 2579–2605.

Weng, S. F.; Reps, J.; Kai, J.; Garibaldi, J. M.; and Qureshi, N. 2017. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PloS one*, 12(4): e0174944.

Wilson, P. W.; D'Agostino, R. B.; Levy, D.; Belanger, A. M.; Silbershatz, H.; and Kannel, W. B. 1998. Prediction of coronary heart disease using risk factor categories. *Circulation*, 97(18): 1837–1847.