Supplementary Materials: Causal Feature Selection from Multiple Omics Data Enhances Machine Learning Performance of Long COVID Early Detection

Sindy Licette Pinero
pinsy007@mymail.unisa.edu.au
University of South Australia
Adelaide, SA, Australia

Thi Nhu Ngoc Duong
duoty016@mymail.unisa.edu.au
University of South Australia
Adelaide, SA, Australia

Xiaomei Li
maisie.zhang@csiro.au
Commonwealth Scientific and Industrial Research Organisation
Marsfield, NSW, Australia

Lin Liu
lin.liu@unisa.edu.au
University of South Australia
Adelaide, SA, Australia

Jiuyong Li
jiuyong.li@unisa.edu.au
University of South Australia
Adelaide, SA, Australia

Sang Hong Lee
hong.lee@unisa.edu.au
University of South Australia
Adelaide, SA, Australia

Thuc Duy Le
thuc.le@unisa.edu.au
University of South Australia
Adelaide, SA, Australia

CCS Concepts

• **Applied computing → Health informatics**; *Bioinformatics*; • **Computing methodologies → Machine learning approaches**; *Causal reasoning and diagnostics.*

Keywords

Long COVID, PASC, early detection, causal inference, differential causal effects, TabPFN, gene expression, machine learning, foundation models, biomarkers

ACM Reference Format:
Sindy Licette Pinero, Thi Nhu Ngoc Duong, Xiaomei Li, Lin Liu, Jiuyong Li, Sang Hong Lee, and Thuc Duy Le. 2025. Supplementary Materials: Causal Feature Selection from Multiple Omics Data Enhances Machine Learning Performance of Long COVID Early Detection. In *Proceedings of Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25).* ACM, Seoul, Republic of Korea , 7 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Classification Metrics Definitions

1.1 Classification Metrics Terminology

The following terminology is used throughout this study for binary classification evaluation:

- **True Positives (TP):** Correctly predicted Long COVID cases
- **True Negatives (TN):** Correctly predicted non-Long COVID cases
- **False Positives (FP):** Incorrectly predicted Long COVID cases (Type I error)
- **False Negatives (FN):** Missed Long COVID cases (Type II error)

1.2 Performance Metrics

1.2.1 *Accuracy.*

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Represents the overall proportion of correct predictions across all classes.

1.2.2 *Precision (Positive Predictive Value).*

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

Measures the proportion of predicted Long COVID cases that are actually positive, indicating model reliability for positive predictions.

1.2.3 *Recall (Sensitivity, True Positive Rate).*

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

Quantifies the proportion of actual Long COVID cases correctly identified, crucial for clinical applications where missing cases has serious consequences.

1.2.4 *F1-Score.*

$$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

Provides the harmonic mean of precision and recall, balancing both metrics for overall model performance assessment [**sokolova2009**].

CIKM '25, Seoul, Republic of Korea

ACM ISBN 978-1-4503-XXXX-X/2025/11
https://doi.org/XXXXXXX.XXXXXXX

1.2.5 ROC AUC (Area Under the Receiver Operating Characteristic Curve).

$$\text{ROC AUC} = \int_0^1 \text{TPR}(FPR^{-1}(t))dt \tag{5}$$

where TPR = $\frac{TP}{TP+FN}$ (True Positive Rate) and FPR = $\frac{FP}{FP+TN}$ (False Positive Rate).

The ROC AUC measures the model's ability to distinguish between classes across all classification thresholds, with values ranging from 0.5 (random) to 1.0 (perfect) [**fawcett2006**].

2 Machine Learning Models: Equations, Purpose and Rationale

2.1 Linear Models

2.1.1 Logistic Regression.

Equation:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^{p} \beta_i x_i)}} \tag{6}$$

where β_0 is the intercept, β_i are the coefficients for features x_i, and p is the number of features.

Purpose: Serves as a fundamental baseline for binary classification, providing interpretable coefficients that indicate the linear relationship between gene expression features and Long COVID risk.

Rationale: Linear models are essential benchmarks in genomics studies due to their interpretability and ability to handle high-dimensional data [**hastie2009**]. They provide insights into whether the relationship between gene expression and Long COVID can be captured through linear combinations.

2.1.2 Ridge Classifier.

Equation:

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{7}$$

where λ is the regularization parameter controlling the strength of L2 penalty.

Purpose: Addresses multicollinearity in gene expression data through L2 regularization, preventing overfitting when dealing with correlated genetic features.

Rationale: Gene expression data often exhibits high correlation between related genes. Ridge regression maintains all features while shrinking coefficients, making it suitable for scenarios where multiple genes may collectively contribute to the outcome [**hoerl1970**].

2.1.3 Linear Discriminant Analysis (LDA).

Equation:

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k \tag{8}$$

where μ_k is the class mean, Σ is the shared covariance matrix, and π_k is the prior probability.

Purpose: Assumes Gaussian distributions for each class and finds linear combinations of features that best separate Long COVID from non-Long COVID cases.

Rationale: LDA provides dimension reduction capabilities while maintaining discriminative power, potentially useful when gene expression patterns follow multivariate normal distributions within each class [**fisher1936**].

2.2 Instance-Based Models

2.2.1 k-Nearest Neighbors (KNN).

Equation:

$$\hat{y} = \text{mode}\{y_i : x_i \in N_k(x)\} \tag{9}$$

where $N_k(x)$ represents the k nearest neighbors of point x based on distance metric $d(x_i, x_j)$.

Purpose: Makes predictions based on similarity to k nearest samples in the feature space, capturing local patterns in gene expression profiles.

Rationale: Genetic similarity often translates to phenotypic similarity. KNN can identify patients with similar gene expression patterns who may share Long COVID outcomes, providing a non-parametric approach to classification [**cover1967**].

2.3 Neural Network Models

2.3.1 Multi-Layer Perceptron (MLP).

Equation:

$$h^{(l)} = \sigma(W^{(l)} h^{(l-1)} + b^{(l)}) \tag{10}$$

$$\hat{y} = \text{softmax}(W^{(L)} h^{(L-1)} + b^{(L)}) \tag{11}$$

where $h^{(l)}$ is the l-th hidden layer, $W^{(l)}$ and $b^{(l)}$ are weights and biases, and σ is the activation function.

Purpose: Captures non-linear relationships between gene expression features through hidden layers, potentially identifying complex gene interaction patterns.

Rationale: Biological systems exhibit complex non-linear interactions. MLP can model these relationships without requiring explicit specification of interaction terms, potentially uncovering hidden patterns in gene expression data [**rumelhart1986**].

2.3.2 Support Vector Classifier (SVC).

Equation:

$$\min_{w,b,\xi} \frac{1}{2} ||w||^2 + C \sum_{i=1}^{n} \xi_i \tag{12}$$

subject to: $y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i$ and $\xi_i \geq 0$

where w is the weight vector, b is the bias, ξ_i are slack variables, C is the regularization parameter, and $\phi(x_i)$ maps input to higher-dimensional space.

Purpose: Finds optimal hyperplanes that separate Long COVID cases with maximum margin, handling non-linear relationships through kernel transformations.

Rationale: SVC is particularly effective for high-dimensional data like gene expression, where the number of features exceeds the

number of samples. The kernel trick allows modeling of complex decision boundaries [**cortes1995**].

2.4 Tree-Based Models

2.4.1 *Random Forest.*

Equation:

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \tag{13}$$

where each tree T_b is trained on a bootstrap sample $\mathcal{D}_b$ with random feature selection at each split.

Purpose: Combines multiple decision trees with bootstrap sampling and random feature selection to reduce overfitting while maintaining predictive power.

Rationale: Tree-based methods naturally handle feature interactions and non-linearities common in biological data. Random Forest provides built-in feature importance measures and handles mixed data types effectively [**breiman2001**].

2.4.2 *Extra Trees (Extremely Randomized Trees).*

Equation:

$$\hat{y} = \frac{1}{M}\sum_{m=1}^{M} T_m(x, \Theta_m) \tag{14}$$

where Θ_m represents random parameters for both feature and threshold selection in tree T_m.

Purpose: Further randomizes the tree construction process by selecting split points randomly, reducing variance compared to Random Forest.

Rationale: Extra randomization can improve generalization, particularly important when dealing with noisy gene expression data where overfitting to specific expression patterns is a concern [**geurts2006**].

2.4.3 *Bagging Classifier.*

Equation:

$$\hat{y} = \text{mode}\{\hat{f}_1(x), \hat{f}_2(x), ..., \hat{f}_B(x)\} \tag{15}$$

where each $\hat{f}_b$ is trained on bootstrap sample $\mathcal{D}_b^* \sim \mathcal{D}$.

Purpose: Implements bootstrap aggregating by training multiple decision trees on different bootstrap samples of the training data and averaging their predictions.

Rationale: Bagging reduces model variance by combining predictions from multiple trees trained on different data subsets. This approach is particularly valuable for gene expression data where individual trees might overfit to specific patterns [**breiman1996**].

2.4.4 *Gradient Boosting.*

Equation:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \tag{16}$$

$$h_m = \arg\min_h \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + h(x_i)) \tag{17}$$

where h_m is the m-th weak learner and γ_m is the step size.

Purpose: Sequentially builds weak learners that correct errors from previous iterations, focusing on difficult-to-classify cases.

Rationale: Boosting methods excel at reducing bias and can identify subtle patterns in data. This is valuable for Long COVID prediction where the biological mechanisms may involve complex, interconnected pathways [**friedman2001**].

2.4.5 *AdaBoost.*

Equation:

$$\alpha_m = \frac{1}{2}\ln\left(\frac{1-\epsilon_m}{\epsilon_m}\right) \tag{18}$$

$$w_i^{(m+1)} = w_i^{(m)} \exp(-\alpha_m y_i h_m(x_i)) \tag{19}$$

where ϵ_m is the weighted error and w_i are sample weights.

Purpose: Adaptively adjusts the importance of training samples, emphasizing misclassified cases in subsequent iterations.

Rationale: AdaBoost can effectively handle class imbalance issues that may exist in Long COVID datasets, ensuring that minority class patterns are adequately learned [**freund1997**].

2.5 Advanced Ensemble Methods

2.5.1 *XGBoost.*

Equation:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{20}$$

where $\Omega(f_k) = \gamma T + \frac{1}{2}\lambda||\omega||^2$ is the regularization term.

Purpose: Implements optimized gradient boosting with advanced regularization techniques and efficient computation.

Rationale: XGBoost has demonstrated superior performance in many genomics applications due to its ability to handle missing values, regularization capabilities, and optimization for predictive accuracy [**chen2016**].

2.5.2 *LightGBM.*

Equation:

$$\text{Gain} = \frac{1}{2}\left[\frac{(\sum g_L)^2}{H_L+\lambda} + \frac{(\sum g_R)^2}{H_R+\lambda} - \frac{(\sum g)^2}{H+\lambda}\right] - \gamma \tag{21}$$

where g and H are gradient and hessian statistics, using leaf-wise tree growth.

Purpose: Provides fast gradient boosting with leaf-wise tree growth and categorical feature optimization.

Rationale: LightGBM's efficiency makes it suitable for large-scale genomics data while maintaining competitive predictive performance through its optimized algorithms [**ke2017**].

2.5.3 *CatBoost.*

Equation:

$$\hat{y}_i^{(k)} = \frac{\sum_{j=1}^{i-1} \mathbb{I}[x_j^{(k)} = x_i^{(k)}] \cdot y_j + \alpha}{\sum_{j=1}^{i-1} \mathbb{I}[x_j^{(k)} = x_i^{(k)}] + 1} \tag{22}$$

implementing ordered target statistics for categorical features.

Purpose: Handles categorical features natively and provides robust performance with minimal hyperparameter tuning.

Rationale: CatBoost's automatic handling of categorical features and built-in regularization makes it valuable for mixed data types common in clinical genomics studies [**prokhorenkova2018**].

2.6 Meta-Ensemble Approaches

2.6.1 *Voting Classifier.*

Equation:

$$\hat{P}(y = c|x) = \frac{1}{M} \sum_{m=1}^{M} P_m(y = c|x) \tag{23}$$

where $P_m(y = c|x)$ is the predicted probability from the m-th base classifier.

Purpose: Combines predictions from multiple diverse algorithms through soft voting, leveraging the strengths of different modeling approaches.

Rationale: Different algorithms capture different aspects of the data. Voting classifiers can provide more robust predictions by aggregating diverse perspectives on the gene expression patterns [**kuncheva2004**].

2.6.2 *Stacking Classifier.*

Equation:

$$\hat{y}_i^{(1)} = f_m(x_i^{train \setminus \mathcal{K}_k}) \text{ for } i \in \mathcal{K}_k \tag{24}$$

$$\hat{y} = g(\hat{y}_1^{(1)}, \hat{y}_2^{(1)}, ..., \hat{y}_M^{(1)}) \tag{25}$$

where g is the meta-learner trained on base model predictions.

Purpose: Uses a meta-learner to optimally combine base model predictions, learning the best way to weight different algorithms.

Rationale: Stacking can achieve superior performance by learning how to combine models rather than using simple averaging, potentially identifying which models perform best for different types of gene expression patterns [**wolpert1992**].

2.7 Foundation Model

2.7.1 *TabPFN (Tabular Prior-Fitted Networks).*

Equation:

$$P(y_{test}|x_{test}, D_{train}) = \text{Transformer}(x_{test}, D_{train}) \tag{26}$$

where $D_{train} = \{(x_i, y_i)\}_{i=1}^{n}$ is the training set and the model learns from context without parameter updates.

Purpose: Employs transformer architecture with in-context learning, leveraging causal reasoning principles embedded in its pre-training process.

Rationale: TabPFN represents a paradigm shift from traditional ML by learning from context rather than parameter optimization. Its causal reasoning capabilities align with our DCE-based feature selection, creating a synergistic framework for causal inference in Long COVID prediction [2].

3 Detailed Model Hyperparameters

3.1 Default Parameter Strategy

Purpose: Using default hyperparameters ensures fair comparison across all models without introducing optimization bias toward any particular algorithm.

Rationale: Our study focuses on evaluating the effectiveness of DCE-based feature selection rather than individual model optimization. Default parameters provide a standardized baseline that reflects how these algorithms perform in typical applications.

3.2 Model-Specific Hyperparameters

3.2.1 *TabPFN.*

- Device: CPU
- N_ensemble_configurations: 32 (default)
- No_preprocess_mode: False (default)

3.2.2 *Linear Models.*

- **Logistic Regression:** max_iter=1000, C=1.0, penalty='l2', solver='lbfgs'
- **Ridge Classifier:** alpha=1.0, solver='auto'
- **Linear Discriminant Analysis:** solver='svd', shrinkage=None

3.2.3 *Tree-Based Models.*

- **Random Forest:** n_estimators=100, max_depth=None, min_samples_split=
- **Extra Trees:** n_estimators=100, max_depth=None, min_samples_split=2
- **Bagging Classifier:** n_estimators=10, max_samples=1.0, max_features=1.0
- **Gradient Boosting:** n_estimators=100, learning_rate=0.1, max_depth=3
- **AdaBoost:** n_estimators=50, learning_rate=1.0

3.2.4 *Advanced Ensemble Methods.*

- **XGBoost:** use_label_encoder=False, eval_metric='logloss', n_estimators=100
- **LightGBM:** verbosity=-1, n_estimators=100, learning_rate=0.1
- **CatBoost:** verbose=0, iterations=1000, learning_rate=0.03

3.2.5 *Neural Networks.*

- **Multi-Layer Perceptron:** max_iter=1000, hidden_layer_sizes=(100,), alpha=0.0001
- **Support Vector Classifier:** C=1.0, kernel='rbf', gamma='scale', probability=True

3.2.6 *Instance-Based Models.*

- **k-Nearest Neighbors:** n_neighbors=5, weights='uniform', algorithm='auto'

3.2.7 *Meta-Ensemble Models.*

- **Voting Classifier:** voting='soft', estimators=[LogisticRegression, RandomForest, XGBoost]
- **Stacking Classifier:** final_estimator=LogisticRegression(), cv=5

4 Cross-Validation and Statistical Analysis

4.1 Stratified K-Fold Cross-Validation

Purpose: Ensures balanced representation of Long COVID cases across all folds while providing robust performance estimates.

Rationale: Stratification is crucial for medical datasets where class imbalance may exist. This approach ensures that each fold contains representative samples from both Long COVID and control groups, providing more reliable performance estimates [**kohavi1995**].

4.2 Performance Aggregation and Reporting

Purpose: Provides robust estimates of model performance by aggregating results across multiple cross-validation folds, quantifying both central tendency and variability.

Rationale: Standard deviation indicates model stability and reliability. Models with high mean performance but high variability may be less trustworthy for clinical applications than models with moderate performance but consistent behavior.

4.3 Causal Feature Selection Implementation

We implemented a comprehensive causal feature selection pipeline integrating three complementary methodologies: Transcriptome-Wide Mendelian Randomization (TWMR), Control Theory (CT), and Differential Causal Effects (DCE). Each method captures distinct causal mechanisms underlying Long COVID development, generating feature sets that reflect genetic susceptibility, network controllability, and pathway-level dysregulation, respectively.

4.3.1 Transcriptome-Wide Mendelian Randomization. For the TWMR analysis, we employed the Mt-Robin method [1] to identify genes whose expression levels causally influence Long COVID susceptibility. This approach leverages genetic variants as instrumental variables to establish causal relationships between gene expression and disease risk.

Our implementation integrated 49 tissue-specific eQTL datasets from GTEx V8 (39,832 genes from 1,000 European individuals) with Long COVID GWAS data (3,018 cases and 1,093,995 controls) [6]. The analysis proceeded through several critical steps:

(1) **LD Matrix Construction**: We calculated linkage disequilibrium using 820,792 SNPs from 836 European individuals, applying an r^2 threshold of 0.5 to ensure genetic instrument independence.
(2) **Instrument Selection**: We implemented multi-criteria filtering to select robust genetic instruments, prioritizing SNPs present across multiple tissues with consistent effect directions and requiring genes to be expressed in at least one tissue.
(3) **Causal Effect Estimation**: We applied reverse regression with weighted random slopes and correlated errors to estimate causal effects while accounting for potential invalid instruments.
(4) **Statistical Validation**: Through bootstrapping that preserved LD structure, we generated null distributions to establish statistical significance (FDR < 0.05).
(5) **Risk Score Calculation**: We computed normalized risk scores (S_{Risk}) from absolute effect sizes ($|\beta_y|$) using min-max scaling for cross-gene comparability.

This analysis identified genes with statistically significant causal effects on Long COVID risk, providing our first set of mechanistically-grounded features.

4.3.2 Control Theory Network Analysis. The CT analysis [8] identified critical regulatory nodes within Long COVID-associated biological networks. We integrated RNA-seq expression profiles from 413 Long COVID samples [5] with the human protein-protein interaction network [7] to contextualize network control within disease-specific biology.

Our CT implementation involved:

(1) **Network Construction**: We mapped disease-specific gene expression data onto the PPI network structure, creating a Long COVID-contextualized interaction network with directed edges representing regulatory relationships.
(2) **Driver Node Analysis**: We systematically removed each node and measured changes in the minimum number of driver nodes required for full network control, identifying indispensable genes whose removal increased control requirements.
(3) **Gene Classification**: We refined indispensable genes into Type-I (affecting other driver nodes) and Type-II (requiring external control, $K_{in} = 0$) categories based on their control-theoretic properties.
(4) **Network Score Calculation**: We computed S_{Network} by weighting each gene's connectivity (total degree K) by its control-theoretic importance, with Type-II critical genes receiving the highest weight (2.0) due to their requirement for external control.

This analysis revealed genes serving as critical control points in Long COVID pathophysiology, providing network-based causal features complementary to the genetic instruments from TWMR.

4.3.3 Differential Causal Effects Analysis. The DCE framework [3] quantified pathway-level causal differences between COVID-19 and Long COVID states, identifying biological processes with differential causal impacts on disease progression.

Our DCE implementation proceeded as follows:

(1) **Pathway Network Construction**: We extracted gene-gene interactions from KEGG pathways [4], focusing on ECrel (enzyme-enzyme relationships) and GErel (gene expression relationships) while excluding PPrel (protein-protein interactions) as they occur post-transcriptionally.
(2) **Causal Effect Estimation**: Using RNA-seq data from COVID-19 and Long COVID patients, we estimated causal coefficients (β) for each source-target gene pair within pathways, quantifying how regulatory relationships differ between disease states.
(3) **Pathway Prioritization**: We retained pathways containing gene pairs with statistically significant absolute causal effects ($|\beta|$ > threshold, FDR < 0.05), identifying biological processes most disrupted in Long COVID.
(4) **Gene Set Extraction**: From top-ranking pathways, we accumulated genes based on absolute beta coefficients, creating feature sets of varying sizes (44, 100, 200, 300, and 411 genes) to evaluate the optimal balance between biological coverage and predictive performance.

This analysis provided pathway-contextualized causal features that capture the differential regulatory dynamics distinguishing Long COVID from acute infection.

4.3.4 Integrated Feature Set Generation. From these three causal analyses, we generated multiple feature sets for systematic evaluation:

- **31CG (31 Causal Genes)**: Highest-scoring genes from integrated MR and CT analysis using $S_{\text{Causal}} = S_{\text{Risk}} + S_{\text{Network}}$
- **44DCEG and 411DCEG**: Differential causal effect genes from pathway analysis
- **500CTG**: Comprehensive control theory genes capturing broader network influences
- **100-300 DCEG-Pathways**: Optimized pathway-based feature sets of varying sizes
- **31CG-411DCEG**: Combined high-risk and pathway-contextual genes

These causally-informed feature sets, each traceable to explicit mechanistic hypotheses, formed the foundation for our comprehensive ML model evaluation demonstrating the superiority of causal over variance-based feature selection.

4.4 Experimental Setup

To comprehensively evaluate the impact of causal feature selection on model performance, we conducted three experiments with progressively optimized configurations. Each experiment evaluated 16 ML models plus the TabPFN foundation model across all feature subsets using stratified 5-fold cross-validation to ensure robust performance estimates.

4.4.1 Experiment 1: Baseline Configuration. The first experiment established baseline performance using default hyperparameters for all models. This configuration employed:

- **Model Settings**: All models initialized with scikit-learn default parameters
- **Preprocessing**: One-hot encoding for categorical variables with first category dropped
- **Standardization**: Applied only to distance-based models (SVC, KNN) via pipelines
- **Ensemble Methods**: Voting classifier with soft voting; stacking with logistic regression meta-learner
- **TabPFN**: Default configuration on CPU device

This baseline experiment provided unbiased performance estimates without hyperparameter tuning, allowing us to assess the raw discriminative power of causal versus non-causal features.

4.4.2 Experiment 2: Moderate Hyperparameter Tuning. The second experiment introduced moderate hyperparameter optimization to improve model performance while maintaining computational efficiency:

- **Tree-based Models**:
  - n_estimators = 100 (RandomForest, GradientBoosting, AdaBoost, Bagging, ExtraTrees)
  - max_depth = None (allowing full tree growth)
  - learning_rate = 0.1 (GradientBoosting, AdaBoost)
- **Linear Models**:
  - C = 1.0 (LogisticRegression, SVC) for standard regularization
  - max_iter = 1000 to ensure convergence
- **Boosting Models**:
  - XGBoost: eval_metric='logloss', same tree parameters
  - LightGBM: verbosity=-1, same tree parameters
  - CatBoost: verbose=0, iterations=100, same learning rate
- **Other Models**: KNN with n_neighbors=5, MLP with max_iter=1000

This configuration balanced model complexity with training efficiency, providing improved performance while avoiding extensive grid search.

4.4.3 Experiment 3: Fully Optimized Configuration. The third experiment implemented comprehensive optimization with class balancing and refined hyperparameters:

- **Class Balancing**: class_weight='balanced' for models supporting it (LogisticRegression, SVC, RandomForest, ExtraTrees, LightGBM)
- **Advanced Tree Parameters**:
  - Maintained n_estimators=100, max_depth=None
  - Early stopping rounds = 50 for gradient boosting methods
- **Neural Network Enhancement**:
  - MLP: hidden_layer_sizes=(100,), max_iter=1000
  - Proper weight initialization with random_state
- **Feature Scaling**: Comprehensive StandardScaler application for all relevant models
- **Model Cloning**: Fresh model instances for each fold using sklearn.base.clone() to prevent data leakage
- **Regularization**:
  - Ridge: alpha=1.0
  - Elastic net components with balanced penalties

This fully optimized setup maximized each model's potential while addressing class imbalance, a critical consideration given the varying prevalence of Long COVID across cohorts.

4.4.4 Evaluation Metrics and Statistical Analysis. Across all experiments, we computed five key metrics:

- **Accuracy**: Overall classification correctness
- **Precision**: Positive predictive value for Long COVID detection
- **Recall**: Sensitivity for identifying Long COVID cases
- **F1 Score**: Harmonic mean balancing precision and recall
- **ROC-AUC**: Area under the receiver operating characteristic curve

Results were aggregated across folds to compute mean performance and standard deviation, enabling assessment of both average performance and stability. Statistical significance of improvements from causal feature selection was evaluated using paired t-tests across fold-wise results.

4.4.5 Implementation Details. All experiments were implemented in Python 3.11 using:

- scikit-learn 1.3.0 for base ML models
- TabPFN 0.1.0 for the foundation model
- XGBoost 1.7.6, LightGBM 4.0.0, CatBoost 1.2 for gradient boosting
- Stratified K-fold cross-validation with random_state=42 for reproducibility

5 Code Availability

The complete implementation of this framework is available at: https://github.com/SindyPin/Early_Detection_LongCOVID

The repository includes comprehensive documentation, implementation scripts, and reproducibility guidelines for the DCE-TabPFN framework.

References

[1] K. J. Gleason, F. Yang, and L. S. Chen. 2021. A robust two-sample transcriptome-wide mendelian randomization method integrating gwas with multi-tissue eqtl summary statistics. *Genetic Epidemiology*, 45, 4, (June 2021), 353–371. doi:10.1002/gepi.22380.

[2] Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. 2023. TabPFN: a transformer that solves small tabular classification problems in a second. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=cp5PvcI6w8_.

[3] K. P. Jablonski, M. Pirkl, D. Ćevid, P. Bühlmann, and N. Beerenwinkel. 2022. Identifying cancer pathway dysregulations using differential causal effects. *Bioinformatics*, 38, 6, 1550–1559. doi:10.1093/bioinformatics/btab847.

[4] M. Kanehisa and S. Goto. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28, 1, 27–30. doi:10.1093/nar/28.1.27.

[5] R. C. Thompson et al. 2023. Molecular states during acute COVID-19 reveal distinct etiologies of long-term sequelae. *Nature Medicine*, 29, 236–246. doi:10.1038/s41591-022-02107-4.

[6] L. Vilma et al. 2023. Genome-Wide Association Study of Long COVID. *medRxiv*. doi:10.1101/2023.06.29.23292056.

[7] A. Vinayagam, U. Stelzl, R. Foulle, S. Plassmann, M. Zenkner, J. Timm, H. E. Assmus, M. A. Andrade-Navarro, and E. E. Wanker. 2011. A directed protein interaction network for investigating intracellular signal transduction. *Science Signal*, 4. doi:10.1126/scisignal.2001699.

[8] A. Vinayagam et al. 2016. Controllability analysis of the directed human protein interaction network identifies disease genes and drug targets. *Proceedings of the National Academy of Sciences of the United States of America*, 113, 18, 4976–4981. doi:10.1073/pnas.1603992113.