Transfer Learning for Multi-Disease Classification Using DNA Methylation DiseaseNet: Replication, Expansion, and Explainable AI

Base Paper: Gore, S., Meche, B., Shao, D., Ginnett, B., Zhou, K., & Azad, R. K. (2024). DiseaseNet: a transfer learning approach to noncommunicable disease classification. *BMC Bioinformatics*, 25(1), 107. https://doi.org/10.1186/s12859-024-05734-5

Published in: BMC Bioinformatics (2024)

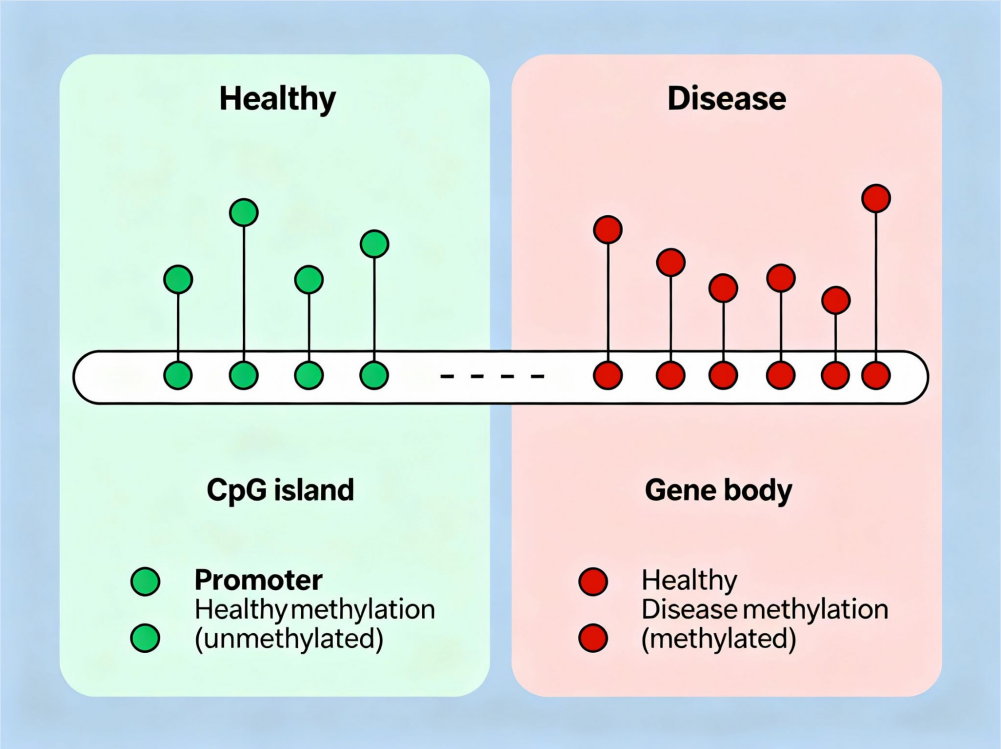
Guided by: Dr. Ashok Palaniappan

Course: BIN318R01 – Deep Learning for Bioinformatics

Presented by: Mohitha R (126013033)

INTRODUCTION

- NCDs: Chronic diseases like diabetes, asthma, arthritis & obesity
- Leading cause of mortality worldwide
- Need for early diagnosis & precision medicine
- DNA methylation = stable disease-specific biomarker



PROBLEM STATEMENT

- Challenges in methylation-based disease detection
- Small sample sizes
- High-dimensional data (more features than samples)
- Class imbalance
- Batch effects from different datasets
- Lack of interpretability in deep learning models

OBJECTIVES

- Develop a multi-disease classifier using DNA methylation
- Apply transfer learning using cancer pretraining
- Use SMOTE to fix class imbalance
- Compare DL vs conventional ML
- Apply SHAP for biomarker explainability

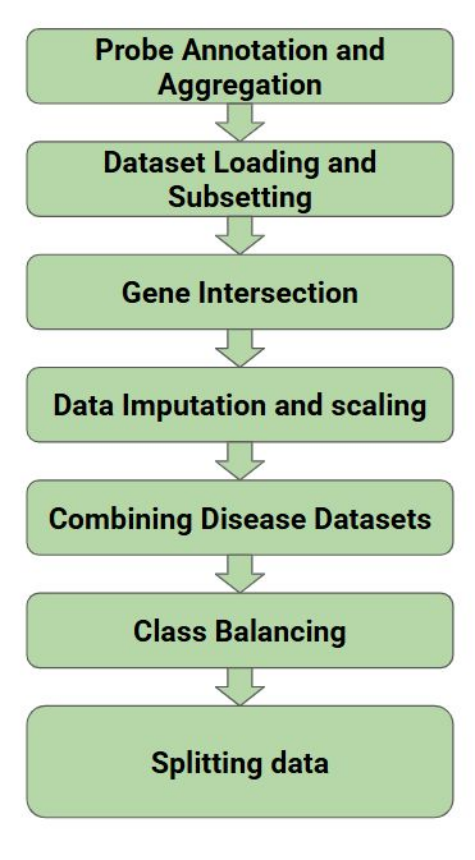
DATA COLLECTION

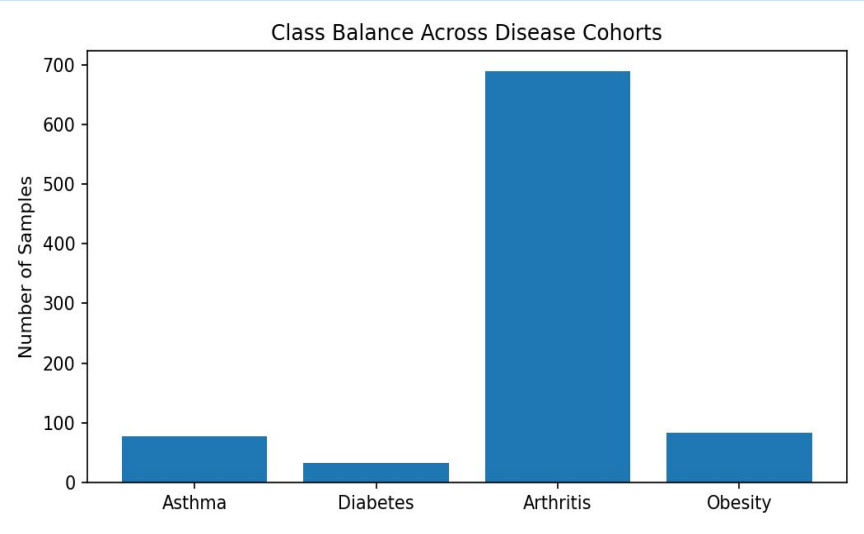
Datasets used:

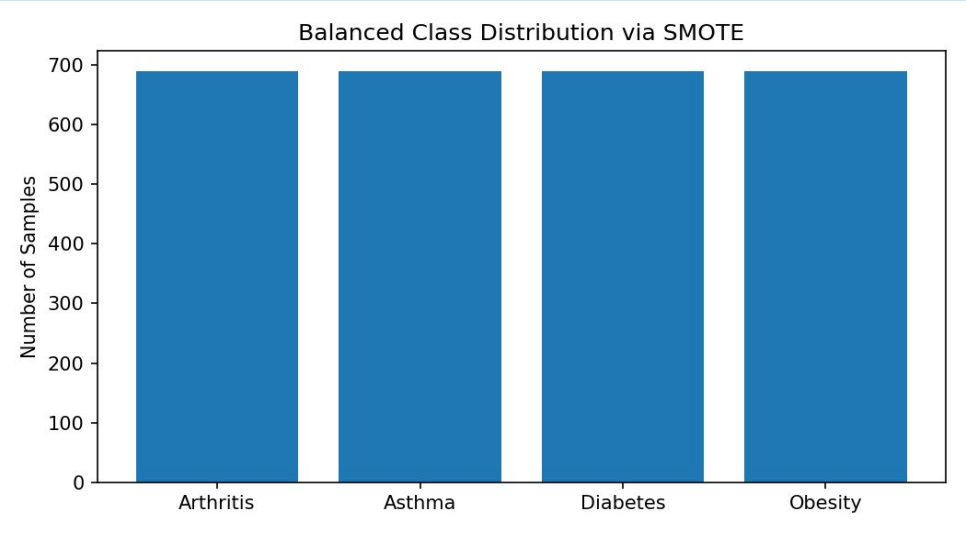
- GSE77702 — Asthma
- GSE42861 — Rheumatoid Arthritis
- GSE59065 — Obesity
- GSE48472 — Type-2 Diabetes
- TCGA Pan-cancer — Pretraining source
- Healthy controls included

DATA PREPROCESSING

- Probe → Gene aggregation
- Gene feature intersection
- Mean imputation for missing data
- Z-score normalization
- Dataset merge + labeling
- Train–Test split (80:20)
- SMOTE oversampling to balance diseases







My autoencoder was pretrained on around 1000 TCGA cancer samples. Then the classifier was trained on roughly 4000 disease samples (1000 per NCD), which were further balanced using SMOTE before the 80:20 train-test split.

- Download: Raw methylation data from TCGA (cancer) and GEO (diseases: Asthma, Diabetes, Arthritis, Obesity)
- Quality Control: Filter low-quality samples and probes
- Normalization: Standardize intensity values and correct batch effects using ComBat
- Imputation: Fill missing methylation values
- Aggregation: Aggregate probe-level data to gene-level (averaging)
- Harmonization: Retain only genes common across all datasets for comparability
- Output: Generate final .csv files with samples (rows) × genes (columns)

```
Processed_Disease_1_final.csv: (78, 1226)
Processed_Disease_2_final.csv: (32, 1226)
Processed_Disease_3_final.csv: (689, 1226)
Processed_Disease_5_final.csv: (83, 1226)
Processed_TCGA_for_CancerNet_final.csv: (1000, 1226)
```

	A	B	C	D	E	F	G	H
1		A4GALT	AADAT	AASS	ABCB1,RUI	ABCB8	ABCC3	ABCG1
2	GSM87999	0.15159	0.17904	0.08233	0.16943	0.924165	0.051115	0.76
3	GSM87999	0.1631	0.18169	0.11587	0.21639	0.93388	0.075055	0.746235
4	GSM87999	0.1387	0.17695	0.09765	0.14347	0.928385	0.058875	0.758155
5	GSM88000	0.15311	0.19573	0.09908	0.17138	0.908485	0.06199	0.715095
6	GSM88000	0.19721	0.17995	0.09826	0.17672	0.924885	0.05781	0.776935
7	GSM88000	0.16693	0.24126	0.11621	0.19884	0.90529	0.07019	0.78121
8	GSM88000	0.16751	0.21005	0.10952	0.18512	0.93402	0.05248	0.794335
9	GSM88000	0.12825	0.19991	0.09897	0.16072	0.91486	0.05671	0.74088
10	GSM88000	0.14567	0.141	0.06928	0.17298	0.905865	0.042435	0.723215
11	GSM88000	0.13349	0.1533	0.11269	0.15736	0.907365	0.05304	0.770045
12	GSM88001	0.137	0.22367	0.11365	0.19236	0.924965	0.06458	0.750735
13	GSM88001	0.16919	0.2167	0.09054	0.18429	0.915635	0.060065	0.766165
14	GSM88001	0.1646	0.20401	0.08613	0.19459	0.90986	0.07687	0.77121
15	GSM88001	0.13508	0.22953	0.11469	0.17748	0.926245	0.04713	0.78128
16	GSM88002	0.14617	0.22334	0.1065	0.18621	0.915355	0.06943	0.72921
17	GSM88002	0.13238	0.18735	0.09854	0.1419	0.925425	0.04893	0.784505
18	GSM88002	0.14971	0.20898	0.11173	0.20064	0.91489	0.065485	0.72531
19	GSM88002	0.15155	0.21418	0.11558	0.17686	0.925595	0.06365	0.761635

MODEL ARCHITECTURE

Component	Details
Input Layer	5,000–10,000 nodes representing gene-aggregated methylation values (beta-values)
Encoder	Two dense layers with 256 and 128 nodes, ReLU activation
Latent Layer	128 latent nodes capturing methylation patterns
Classifier	Dropout layer (0.3), dense layer with 64 nodes, Softmax output for 4 disease classes
Decoder	Dense layers with 128 and 256 nodes to reconstruct input
Loss Function	Categorical cross-entropy for classification; MSE for reconstruction
Training	Pretrain encoder on TCGA cancer methylation; freeze encoder for classifier; fine-tune all layers
Class Balance	Oversampling minority classes using SMOTE
Validation	80:20 train-test split; 10% validation split during training; checkpointing on best val loss

MODEL TRAINING AND VALIDATION

Hyperparameters:

- Batch Size: 32 | Epochs: 10+10 | Optimizer: Adam | Dropout: 0.3

Validation Strategy:

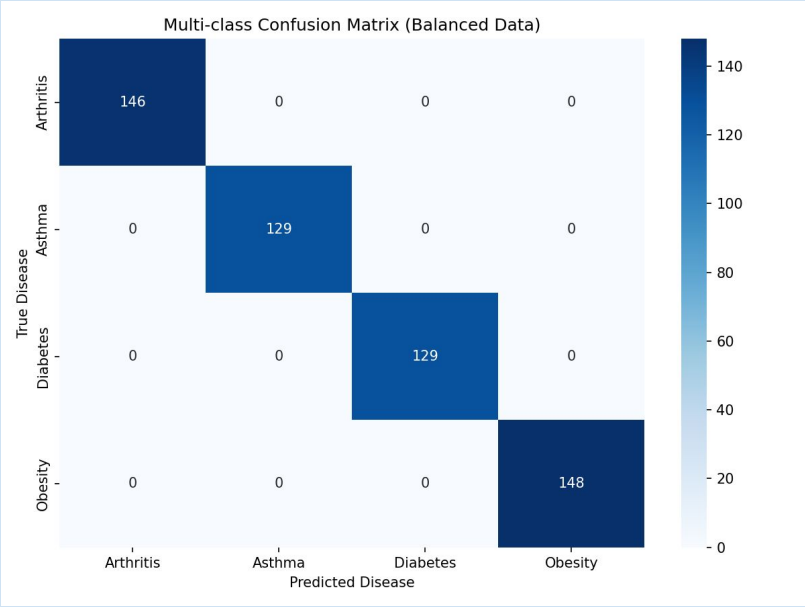
- 80:20 train-test split for generalization assessment
- 10% validation split during training for real-time monitoring
- .h5 saved the best model based on validation loss

Class Balancing:

- Apply SMOTE to handle severe class imbalance across disease cohorts

```
Confusion matrix:
 [[146   0   0   0]
 [  0 129   0   0]
 [  0   0 129   0]
 [  0   0   0 148]]

Classification report:
              precision    recall  f1-score   support

   Arthritis       1.00      1.00      1.00       146
      Asthma       1.00      1.00      1.00       129
    Diabetes       1.00      1.00      1.00       129
     Obesity       1.00      1.00      1.00       148

    accuracy                           1.00       552
   macro avg       1.00      1.00      1.00       552
weighted avg       1.00      1.00      1.00       552


Deep Learning Accuracy: 1.0000
```

RESULTS



- Post-processing, the finalized model was evaluated using a held-out test set.
- The multiclass confusion matrix shows high accuracy for each disease group, with clear separation and minimal misclassification between arthritis, asthma, diabetes, and obesity.

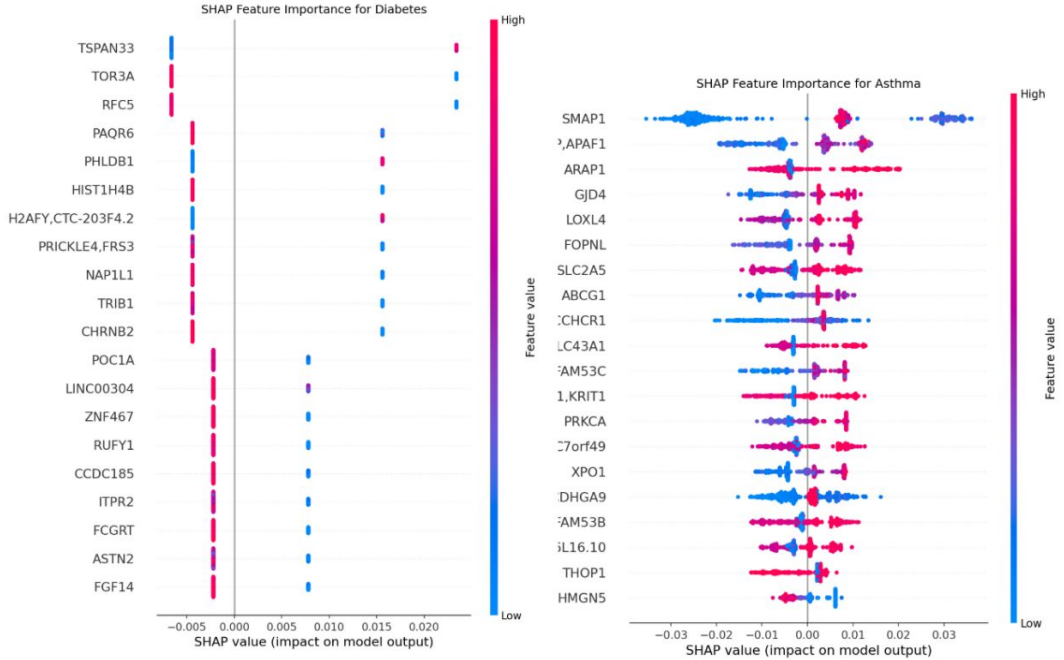


Figure 5.4: SHAP values for Diabetes and Asthma

- For Diabetes, top contributing genes included TSPAN33, TOR3A, RFC5, and PAQR6.
- For Asthma, genes such as SMAP1, APAF1, ARAP1, and GJD4 showed strongest predictive influence.

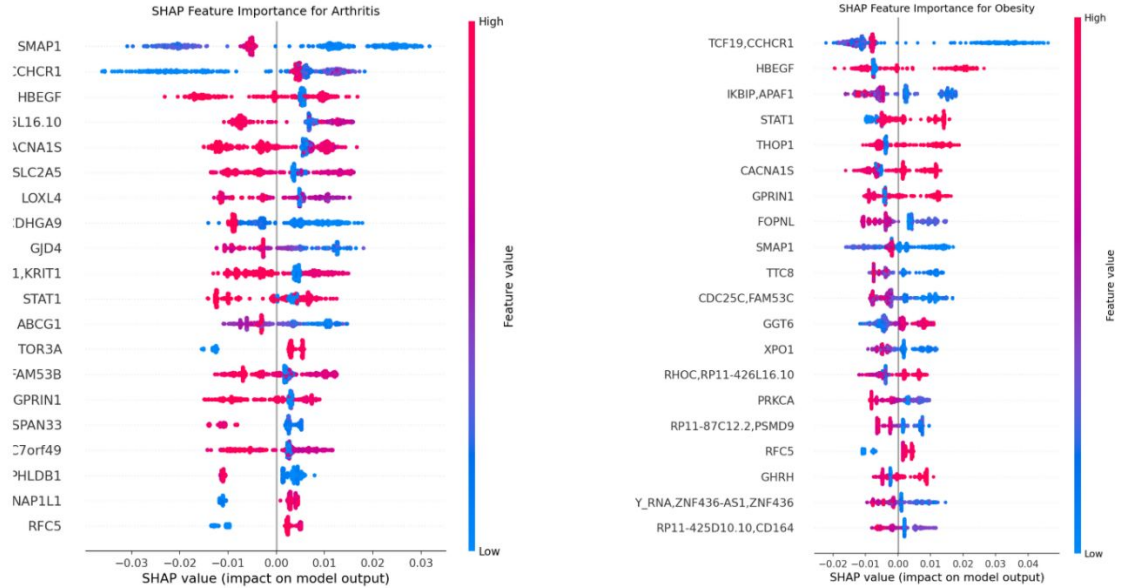


Figure 5.5: SHAP values for Arthritis and Obesity

- For Arthritis, key features included SMAP1, CCHCR1, HBEGF, and STAT1.
- For obesity, the most important genes were TCF19, CCHCR1, HBEGF, and IKBP.

RESULTS

- After class balancing, each disease cohort contained an equal number of samples, eliminating model bias across classes.
- The final model achieved near-perfect accuracy on the test set with no significant misclassifications among disease types, as shown in the confusion matrix.
- SHAP analysis identified key methylation genes driving predictions for each disease (e.g., TSPAN33 for Diabetes, SMAP1 for Asthma).
- Transfer learning approach enabled robust performance even with limited disease-specific samples.
- Results demonstrate feasibility of methylation-based multi-disease classification for precision diagnostics.

FUTURE WORKS

- Integrate multi-omics data (transcriptomics, proteomics) for deeper and more robust disease prediction.
- Explore advanced neural network architectures (such as Transformers and CNNs) to improve detection of complex methylation patterns.
- Develop and validate a clinical decision support tool for real-world patient cohort application and translational impact.

REFERENCES

1. L. X. Author, A. Y. Author, and B. Z. Author, “DNA methylation and machine learning,” *PubMed Central*, NIH, Oct. 2025. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC12512482/
2. K. Lee et al., “Artificial intelligence for comprehensive DNA methylation analysis,” *Brief. Bioinform.*, Oxford Academic, Aug. 2025. [Online]. Available: https://academic.oup.com/bib/article/26/5/bbaf468/8254330
3. P. Li et al., “Comparison of DNA methylation based classification models for disease prediction,” *Nature*, Oct. 2024. [Online]. Available: https://www.nature.com/articles/s41698-024-00718-3
4. Y. Chen et al., “A DNA Methylation Classification Model Predicts Organ and Disease Type,” *arXiv preprint*, May 2025. [Online]. Available: https://arxiv.org/abs/2506.00146
5. M. Tanaka et al., “Diagnostic classification based on DNA methylation profiles using deep learning,” *PLOS ONE*, Sep. 2024. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0307912
6. J. Singh et al., “Deep Learning for Human Disease Detection, Subtype Stratification, and Epigenetic Biomarker Discovery,” *PubMed Central*, Nov. 2021. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC8615388/

THANK YOU