

DriftNet: A Contrastive Learning for Pan-Cancer Stage Classification

by

**Md Abu Ammar (222014056)
Evana Tanzim Rupa (222014046)
Faria Islam (222014074)
Jannatul Mawa Mim (222014066)**

*Capstone project report (CSE 4098C) submitted in partial fulfillment of the
requirements for the degree of*

Bachelor of Science in Computer Science and Engineering

Under the supervision of

Md. Nazmul Abdal



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

UNIVERSITY OF LIBERAL ARTS BANGLADESH

SPRING 2026

© *Md Abu Ammar, Evana Tanzim Rupa, Faria Islam , and Jannatul Mawwa Mim*
All rights reserved

DECLARATION

Project Title DriftNet: A Contrastive Learning for Pan-Cancer Stage Classification
Authors Md Abu Ammar, Evana Tanzim Rupa, Faria Islam , and Jannatul
Mawa Mim
Student IDs 222014056, 222014046, 222014074, and 222014066
Supervisor Md. Nazmul Abdal

We declare that this capstone project report entitled *DriftNet: A Contrastive Learning for Pan-Cancer Stage Classification* is the result of our own work except as cited in the references. The capstone project report has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Md Abu Ammar
222014056

Department of Computer Science and Engineering
University of Liberal Arts Bangladesh

Evana Tanzim Rupa
222014046

Department of Computer Science and Engineering
University of Liberal Arts Bangladesh

Faria Islam
222014074

Department of Computer Science and Engineering
University of Liberal Arts Bangladesh

Jannatul Mawa Mim
222014066

Department of Computer Science and Engineering
University of Liberal Arts Bangladesh

Date: May 03, 2026



Department of Computer Science and Engineering
University of Liberal Arts Bangladesh
Mohammadpur, Dhaka - 1207

CERTIFICATE

This is to certify that the capstone project report entitled **DriftNet: A Contrastive Learning for Pan-Cancer Stage Classification**, submitted by **Md Abu Ammar** (Student ID: 222014056), **Evana Tanzim Rupa** (Student ID: 222014046), **Faria Islam** (Student ID: 222014074) and **Jannatul Mawa Mim** (222014066) are undergraduate students of the **Department of Computer Science and Engineering** has been examined. Upon recommendation by the examination committee, we hereby accord our approval of it as the presented work and submitted report fulfill the requirements for its acceptance in partial fulfillment for the degree of *Bachelor of Science in Computer Science and Engineering*.

Md. Nazmul Abdal
Lecturer

Department of Computer Science and Engineering
University of Liberal Arts Bangladesh

Foysal Ahmed
Lecturer

Department of Computer Science and Engineering
University of Liberal Arts Bangladesh

Fatema Khan
Lecturer

Department of Computer Science and Engineering
University of Liberal Arts Bangladesh

Place: Dhaka

Date: May 03, 2026

ACKNOWLEDGEMENTS

We would like to express our deep and sincere gratitude to our research supervisor, *Md. Nazmul Abdal*, for giving us the opportunity to conduct research and providing invaluable guidance throughout this work. His dynamism, vision, sincerity and motivation have deeply inspired us. He has taught us the methodology to carry out the work and to present the works as clearly as possible. It was a great privilege and honor to work and study under his guidance.

We are greatly indebted to our honorable teachers of the Department of Computer Science and Engineering at the University of Liberal Arts Bangladesh who taught us during the course of our study. Without any doubt, their teaching and guidance have completely transformed us to the persons that we are today.

We are extremely thankful to our parents for their unconditional love, endless prayers and caring, and immense sacrifices for educating and preparing us for our future. We would like to say thanks to our friends and relatives for their kind support and care.

Finally, we would like to thank all the people who have supported us to complete the project work directly or indirectly.

Md Abu Ammar, Evana Tanzim Rupa, Faria Islam , and Jannatul Mawa Mim

University of Liberal Arts Bangladesh

Date: May 03, 2026

Dedicated to
*Almighty Allah, whose mercy
turned every hardship into
strength and every effort into
meaning. I also dedicate this
work to my beloved family
and wonderful teammates,
whose prayers, sacrifices,
support, and faith became the
light behind this achievement.*

– Md Abu Ammar

Dedicated to
*my sister, parents, and
wonderful teammates for
their endless support,
strength, and faith.*

– Evana Tanzim Rupa

Dedicated to
*my family, my greatest
blessing.*

– Faria Islam

Dedicated to
*Almighty Allah for
everything I have in my life
and for making everything
easier.*

– Jannatul Ma'wa Mim

ABSTRACT

Identifying the stage of cancer is crucial because it greatly helps the chances of successful treatment. In this study, we developed a new intelligent system, named DriftNet, for the early detection of cancer stages. First, large DNA methylation data are compressed into smaller meaningful signals using an autoencoder. Then, we trained the model adopting contrastive learning to better distinguish cancer stages, which pulls similar cases together and pushes apart different ones. Finally, we combined these with simple patient information such as age and cancer type in a two-branch deep learning model. DriftNet is tested on the TCGA dataset, from 1,343 patients across multiple cancer-types. In the final evaluation, DriftNet achieved the highest test accuracy of 67.82%, the highest macro F1-score of 0.6442, and the highest Stage III F1-score of 0.5138. It outperformed both classical machine-learning baselines and paper-inspired deep-learning baselines, especially for the difficult Stage III class. These results show that supervised contrastive learning and dual-branch fusion can improve pan-cancer Stage I/II/III classification using methylation and clinical information.

Keywords: Cancer staging, Contrastive learning, Deep learning, DNA methylation, Pan-cancer classification, TCGA, autoencoder

Contents

1	Introduction	1
1.1	Problem statement	2
1.2	Aims, objectives & Motivation	2
1.2.1	Aims	2
1.2.2	Objectives	2
1.2.3	Motivation	3
1.3	Project Specification	3
1.4	Expected Outcomes	4
2	Literature Review	5
2.1	DNA Methylation in Cancer Research	5
2.2	TCGA as a Pan-Cancer Resource	6
2.3	Dimensionality Reduction and Autoencoders	6
2.4	Contrastive Learning and Triplet-Loss Embeddings	7
2.5	Deep Learning and Ensemble Methods for Cancer Prediction	8
2.6	Handling Class Imbalance with Focal and Related Losses	9
2.7	Multi-Modal and Multi-Omics Integration	10
2.8	Evaluation Practices and Stability Analyses	11
2.9	Positioning DriftNet	11
3	Methodology	13
3.1	Autoencoder-Based Methylation Representation	14
3.2	Clinical Feature Encoding	15
3.3	Triplet-Based Contrastive Feature Learning	16
3.4	DriftNet Stage Classification Model	16
3.5	Baseline and Paper-Based Validation Models	17

4	Experiments and Result Analysis	20
4.1	Dataset and Pre-processing	20
4.2	Experimental Setting	22
4.3	Evaluation Metrics	23
4.4	Analysis of Experimental Results	24
4.4.1	Autoencoder Compression and Latent Dimension Selection .	24
4.4.2	Contrastive Learning Improves Stage Separability	25
4.4.3	DriftNet Stage Classification Performance	27
4.4.4	Cross-Validation Analysis	29
4.4.5	Comparison with Classical and Paper-Based Baseline Models	30
4.4.6	Impact of Dimensionality Reduction: Comparing Full dimen- sional and Autoencoder-based Latent Space	32
4.4.7	Error Analysis and Misclassification Patterns	33
5	Design and Development	35
5.1	Design Overview	35
5.2	Web Interface Design	36
5.2.1	Inference Workflow View	36
5.2.2	Main Prediction Page Layout	37
5.3	Major Subsystems in the Design	38
5.3.1	Frontend Interface Subsystem	38
5.3.2	Backend Processing Subsystem	39
5.3.3	Machine-Learning Inference Subsystem	39
5.4	Input Design	40
5.5	Output Design	41
5.6	Engineering Analysis and Design Decisions	42
5.6.1	Input Validation	42
5.6.2	Feature Dimension Control	42
5.6.3	Pipeline Consistency	42
5.6.4	Usability and Error Reduction	42
5.7	Discussion of the Developed Design	43
5.8	Development Outcome	43

6	Plans and Deliverables	44
6.1	Timeline	44
6.2	Budget Planning	45
7	Conclusion	46
7.1	Social, Legal, Ethical, and Environmental Issues	48
7.2	Expected Final Product	48
	References	50

List of Figures

3.1	Overall DriftNet methodology workflow.	14
4.1	Autoencoder reconstruction loss across latent dimensions.	25
4.2	A side-by-side t-SNE plot.	27
4.3	DriftNet Performance Dashboard.	28
4.4	ROC curves and learning curve of the best DriftNet model.	29
4.5	Five-fold cross-validation performance summary.	30
5.1	DriftNet homepage interface.	36
5.2	Four-step DriftNet inference workflow.	37
5.3	Main DriftNet prediction interface.	38
5.4	DriftNet input interface after data entry.	40
5.5	DriftNet prediction output interface.	41
6.1	Project timeline for the 12-month DriftNet research study	44

List of Tables

4.1	Final cohort distribution by diagnosis and stage.	21
4.2	Input features used for modelling	22
4.3	DriftNet performance on the held-out test set.	28
4.4	Five-fold stratified cross-validation results.	30
4.5	Final comparison of DriftNet with comparison models.	31
4.6	Performance comparison of full-dimensional and autoencoder features.	33
4.7	DriftNet misclassification patterns on the test set.	34
6.1	Estimated budget planning for the 12-month DriftNet research project	45

Chapter 1

Introduction

Cancer staging is essential for treatment plans, as it describes the size of the tumour and the extent of spread in the body, guiding decisions [Edge et al. \(2010\)](#). Incorrect staging may cause insufficient or excessive treatment, impacting important parameters, such as survival and quality of life [Brierley et al. \(2017\)](#). Clinical staging usually depends on imaging, operative assessment, and microscopic evaluation of tissue; however, these require expert review and may still miss subtle molecular changes associated with early or progressing disease [Weisenberger et al. \(2014\)](#); [Hao et al. \(2017\)](#). These limitations motivate the development of intelligent methods that complement and refine clinical staging by directly capturing tumour biology [Ivan et al. \(2021\)](#).

In this study, we propose a deep-learning system that uses DNA methylation to perform pan-cancer staging in multiple types of tumour. DNA methylation changes in a systematic manner as tumours start, grow, and spread [Hao et al. \(2017\)](#), and large projects such as TCGA supply genome-wide methylation profiles with matched clinical staging data [Weinstein et al. \(2013\)](#). However, methylation arrays contain more than 450000 CpG sites per sample, and classical models struggle with such high-dimensional and noisy features [Guyon & Elisseeff \(2003\)](#). In light of this, our approach first encodes genome-scale methylation using a variational autoencoder [Gomari et al. \(2022\)](#). It then applies supervised contrastive learning with triplet loss to learn stage-aware representations from the VAE-derived methylation features and encoded clinical variables [Schroff et al. \(2015\)](#). Finally, these learned representations are integrated with clinical information in a dual-branch neural network for Stage I–III classification [Bengio et al. \(2013\)](#).

1.1 Problem statement

Clinical staging usually depends on imaging, operative assessment, and microscopic evaluation of tissue; however, these require expert review and may still miss subtle molecular changes associated with early or progressing disease [Weisenberger et al. \(2014\)](#)[Hao et al. \(2017\)](#). Incorrect staging may cause insufficient or excessive treatment, impacting important parameters, such as survival and quality of life [Brierley et al. \(2017\)](#).

Although DNA methylation changes in a systematic manner as tumours start, grow, and spread [Hao et al. \(2017\)](#), methylation arrays contain more than 450000 CpG sites per sample, and classical models struggle with such high-dimensional and noisy features [Guyon & Elisseeff \(2003\)](#). As a result, there is a need for intelligent methods that can effectively compress, interpret, and utilize genome-scale methylation data to improve cancer staging accuracy.

1.2 Aims, objectives & Motivation

The limitations of traditional staging methods and the availability of large-scale methylation datasets motivate the development of deep-learning frameworks that can capture tumour biology more directly. This study addresses this need by proposing a multi-stage deep-learning system that integrates molecular and clinical information for pan-cancer staging.

1.2.1 Aims

The main aim of this study is to propose a deep-learning system that uses DNA methylation to perform pan-cancer staging in multiple types of tumour by integrating genome-scale methylation features with basic clinical variables.

1.2.2 Objectives

The objectives of this study are:

- To encode genome-scale DNA methylation data using variational autoencoders to compress more than 450000 CpG features into a compact latent representation.

- To apply supervised contrastive learning with triplet loss to enhance the separability of cancer stages, particularly between Stage II and Stage III.
- To design a dual-branch neural network that integrates contrastive stage-aware embeddings with clinical variables for final Stage I–III classification.
- To compare DriftNet with both classical machine-learning baselines and paper-inspired deep-learning baselines, including multimodal neural fusion, DBN-ELM-BP-inspired learning, autoencoder-based ANN classification, and OmiVAE-style VAE classification.

1.2.3 Motivation

These limitations motivate the development of intelligent methods that complement and refine clinical staging by directly capturing tumour biology [Ivan et al. \(2021\)](#). Large projects such as TCGA supply genome-wide methylation profiles with matched clinical staging data [Weinstein et al. \(2013\)](#), creating an opportunity to build robust pan-cancer staging models. By combining variational autoencoders, supervised contrastive learning, and deep neural networks, this study is motivated by the goal of improving stage separability, classification performance, and robustness across heterogeneous cancer types.

1.3 Project Specification

This project aims to build a deep-learning system that can automatically predict the stage of cancer using DNA methylation data and some basic clinical information. The system will use one model to work with multiple cancer types, instead of creating separate models for each disease.

The model should be able to classify cancer cases into three stages: Stage I, Stage II, and Stage III. The success of the model will be measured using simple performance metrics such as accuracy, precision, recall, and F1-score. These metrics will help show whether the model is making correct and reliable predictions.

To improve prediction quality, the system uses a contrastive learning approach so that samples from the same stage become closer in the learned feature space, while samples from different stages become more separated. The improvement

of feature quality will be measured using standard clustering metrics, such as Silhouette Score and Davies–Bouldin Index.

The final project will include the trained model, the full implementation code, and a simple web-based interface. The web interface will allow users to enter patient data and see the predicted cancer stage in an easy-to-understand way.

Project requirements may change slightly during development as testing and evaluation continue. Any changes will be made carefully to ensure that the main goal of building a reliable and usable pan-cancer staging system remains unchanged.

1.4 Expected Outcomes

The expected outcome of this research is an evaluated deep-learning framework called DriftNet, designed to automatically predict cancer stage using DNA methylation data combined with basic clinical information. The model is built as a pan-cancer system, meaning it can be applied to multiple cancer types using a single unified framework.

The final product will be provided as a reproducible research model, including trained networks, source code, and documented results. DriftNet is designed to be extensible, allowing future studies to incorporate additional molecular data or evaluate the model on independent datasets.

The primary beneficiaries of this work include the biomedical research and healthcare sector, particularly researchers and clinicians involved in cancer diagnosis and data-driven decision making. With further validation, the system may support oncologists by assisting in cancer staging, especially in complex cases.

The findings of this project are expected to be shared through academic channels, including thesis submission, potential conference presentation, and journal publication in the field of medical artificial intelligence.

Chapter 2

Literature Review

In this section, we review recent work related to DNA methylation analysis, deep learning, and pan-cancer modeling. We aim to highlight what is currently known, identify the limitations of existing methods, and position DriftNet within this landscape. We organize the discussion into several themes, including methylation-driven cancer research, TCGA-based pan-cancer resources, dimensionality-reduction strategies, contrastive learning frameworks, and deep-learning approaches used in clinical prediction.

2.1 DNA Methylation in Cancer Research

DNA methylation is a stable epigenetic mark that controls gene activity without actually altering the DNA sequence [Weisenberger et al. \(2014\)](#). Many studies have demonstrated that the methylation patterns of cancer cells are different from those of normal cells and that these patterns reflect important information about tumour type, behaviour, and outcome [Hao et al. \(2017\)](#). Large-scale pan-cancer analyses of the TCGA data have mapped common and cancer-specific methylation changes among multiple tumour types and used them in tasks such as tumour-of-origin prediction and detection of aberrant regulatory regions [Hoadley et al. \(2018\)](#). These works prove that methylation is a powerful signal for cancer biology [Esteller \(2008\)](#).

However, the majority methylation-based models are designed for cancer vs normal detection, tumour-of-origin classification, or discovery of subtypes rather than direct clinical stage prediction [Baylin & Jones \(2011\)](#). Some studies have reported stage-specific methylation changes in single cancers such as breast or colorectal cancer, but they usually build separate models for each cancer and for a

small number of markers. There is still no single methylation-driven system that predicts AJCC Stages I, II, and III together in many cancers.

2.2 TCGA as a Pan-Cancer Resource

The Cancer Genome Atlas (TCGA) is one of the most important resources for cancer genomics. It provides matched molecular data, including DNA methylation, and detailed clinical information on age, stage, and outcome for over 11,000 patients across 33 cancer study types [Weinstein et al. \(2013\)](#). TCGA Pan-Cancer Atlas projects have investigated cross-cancer patterns of copy-number changes, expression, and methylation and generated reference maps for many of the hallmarks of cancer [Hoadley et al. \(2018\)](#). Although TCGA contains 33 official cancer study types, most individual patient records are actually classified with much more granular primary diagnosis codes that differentiate anatomical sites and histological subtypes. Because TCGA contains many primary diagnoses measured on comparable platforms, it is useful for pan-cancer machine-learning models [Irizarry et al. \(2009\)](#).

Simultaneously, TCGA methylation data is extremely high-dimensional, with each sample having more than 450,000 CpG probes. This makes the direct use of raw methylation in standard classifiers very difficult, both statistically and computationally [Widschwendter & Jones \(2002\)](#). Many TCGA-based models handle this with heavy feature selection, focusing on small gene panels, or working with single cancer cohorts. A general pan-cancer staging model must, therefore face both the biological heterogeneity across cancers and the technical challenge of ultra-high-dimensional methylation data.

2.3 Dimensionality Reduction and Autoencoders

Dimensionality reduction in omics data has traditionally been performed using classical approaches such as principal component analysis (PCA) or filter-based feature selection [Guyon & Elisseeff \(2003\)](#) [Cruz & Wishart \(2006\)](#). These methods can be helpful but are usually linear and lack the complexity of nonlinear relationships that might exist [Frid-Adar et al. \(2018\)](#). VAEs present a deep learning approach that can model nonlinear structure by learning a low-dimensional latent representation and reconstructing the original input [Wei & Mahmood \(2020\)](#). VAEs have been

used to summarize single-cell RNA-seq, metabolomics, and other high-dimensional biological data, and their latent spaces can reveal meaningful biological gradients [Lopez et al. \(2018\)](#).

In cancer genomics, VAEs have commonly been used for multi-omics integration and for learning compact codes to aid downstream prediction tasks [Zhang et al. \(2019\)](#). Yet, a standard VAE’s objective is to minimise reconstruction error plus a regularization term, not separate out specific clinical groups like stages. Consequently, the latent features may be good at capturing variation in general but not optimized for distinguishing Stages I, II, and III. This motivates our use of VAEs only as a first compression step, followed by a second step reshaping the latent space for the purposes of staging specifically.

Recent autoencoder-based studies have also shown that compressed multi-omics features can support cancer classification tasks. Munquad et al. proposed cancer-associated multi-omics latent variables (CMLV), where gene expression, microRNA, and DNA methylation were integrated using an autoencoder [Munquad et al. \(2025\)](#). The latent features from the bottleneck layer were then used with an artificial neural network for tissue-of-origin, subtype, and stage-related classification. This study is relevant to DriftNet because both methods use compressed omics representations before classification. However, CMLV-AE + ANN mainly relies on autoencoder-derived latent features, while DriftNet further applies supervised contrastive learning to make Stage I, Stage II, and Stage III more separable.

OmiVAE is another important VAE-based model for multi-omics cancer classification. It combines autoencoder-based latent feature learning with a classification network [Zhang et al. \(2019\)](#). The model first learns a compact representation from high-dimensional omics data and then uses this representation for multi-class cancer classification. OmiVAE is useful as a baseline for DriftNet because it tests whether VAE-based representation learning alone is enough for classification. In contrast, DriftNet uses VAE only as the first compression step and then improves the representation using triplet-loss contrastive learning.

2.4 Contrastive Learning and Triplet-Loss Embeddings

Contrastive learning refers to a family of methods that learn representations by contrasting pairs or triplets of samples. Triplet loss, originally popularised in

FaceNet for face recognition, takes an anchor sample, a positive sample from the same class, and a negative sample from a different class, and encourages the model to move the anchor closer to the positive and away from the negative in embedding space [Schroff et al. \(2015\)](#). This simple idea has been extremely successful in many computer vision tasks and has been adapted to various medical imaging, such as brain tumour classification and breast lesion segmentation [Liu & Feng \(2021\)](#) [Wu et al. \(2025\)](#).

More recently, it has been applied to multi-modal and multi-omics data, where it aligns different data types, hence reducing redundancy [Zhao et al. \(2023\)](#). Some works have developed their contrastive objectives to improve survival prediction or clustering tasks in multi-omics cohorts [Z. Huang et al. \(2019\)](#). However, to our best knowledge, no earlier work has applied the supervised triplet-loss contrastive learning on VAE-compressed DNA methylation with the express aim of distinguishing cancer stages or combined those embeddings with clinical variables in a pan-cancer Stage I–III classifier. This is the central methodological gap that our work addresses.

2.5 Deep Learning and Ensemble Methods for Cancer Prediction

Deep learning has changed the face of many areas within cancer research. Neural networks and convolutional models have achieved very strong results in various image analysis tasks (pathology, radiology) and in some genomic analyzes, particularly when regularization techniques such as batch normalization, dropout, and early stopping were applied to control overfitting of the models [LeCun et al. \(2015\)](#) [Ioffe & Szegedy \(2015\)](#) [Srivastava et al. \(2014\)](#) [Prechelt \(2002\)](#). Meanwhile, ensemble tree-based models like random forests and gradient boosting retain much of their competitive edge in performance for tabular biomedical data. For instance, extreme gradient boosting (XGBoost) applied with multi-omics features on the classification of various cancers and reported remarkable diagnostic performance. Support vector machines (SVMs) are also widely used in cancer genomics because they perform well in high-dimensional spaces [S. Huang et al. \(2018\)](#).

However, most of these models are designed for a single task like detection, subtype classification, or survival and usually deal with one cancer type at a time during the training process [Kourou et al. \(2015\)](#). Only a few works attempt a unified

pan-cancer model, and even fewer tackle the specific challenge of AJCC staging Powers (2020). In fact, the standard classical baselines, such as logistic regression, SVM, random forest, gradient boosting, and k-nearest neighbours, usually base their operation on hand-engineered or selected features, not on contrastive-learned latent representations. In this work, we consider these baselines, trained on the same feature set, to show the added value of the DriftNet architecture and triplet-based representation learning.

Amanzholova and Coşkun proposed a hybrid deep-learning approach for cancer stage prediction using TCGA multi-omics data Amanzholova & Coşkun (2024). Their method combined deep belief networks, extreme learning machines, and backpropagation to predict early-stage and late-stage cancers. This work is important because it directly focuses on cancer stage prediction using omics and clinical information. However, their stage task mainly separates cancer into early and late groups, while DriftNet directly predicts Stage I, Stage II, and Stage III as three separate classes. For this reason, a DBN-ELM-BP-inspired baseline was included in this study for fair comparison.

2.6 Handling Class Imbalance with Focal and Related Losses

Clinical datasets often exhibit imbalanced class distributions, with early-stage cases being more frequent than late-stage ones, or vice versa, depending on the screening practices Sokolova & Lapalme (2009). The model is then biased toward the majority class by the standard cross-entropy loss, leading to poor recall for the minority classes Lin et al. (2017). Focal loss was proposed as its solution, down-weighting easy, well-classified examples and focusing the training on hard, misclassified ones. Follow-up work adapted focal loss to medical image classification and other challenging imbalanced settings, reporting better performance compared with plain cross-entropy Yeung et al. (2022).

In pan-cancer staging, stage III cases are usually much rarer and more heterogeneous compared to stage I or II. If this imbalance is not treated with much care, the model can overpredict the more common stages and fail to recognize advanced

disease [Gurcan & Soylu \(2024\)](#). Our work employs a simple focal-loss variant together with class weights in order to ensure meaningful contributions of all stages, in particular Stage III, to the training objective.

2.7 Multi-Modal and Multi-Omics Integration

A growing body of work has shown that the integration of different data types, such as DNA methylation, gene expression, copy-number alteration, and clinical features, will provide an enhanced prediction and subtyping of cancer [Rappoport & Shamir \(2018\)](#)[Z. Huang et al. \(2019\)](#). Integration strategies span early fusion (concatenating all features before the model) and late fusion (merging the outputs of separate models) up to more complicated attention-based or graph-based fusion schemes [Finn et al. \(2017\)](#) [Pan & Yang \(2009\)](#). Multi-omics models have been utilised to define robust molecular subtypes and predict prognosis more accurately compared to single-omics models [Cheerla & Gevaert \(2019\)](#).

However, even in multi-modal settings, most models continue to use raw or, at most, simply reduced features. They rarely use contrastive learning to shape the representations of each modality before fusing them, and few treat DNA methylation as a primary signal for staging across cancers [Ribeiro et al. \(2016\)](#). Our design of DriftNet accordingly follows a dual-path architecture: Our design of DriftNet follows a dual-path architecture: one path processes the 128-dimensional contrastive stage-aware embedding, and the other path processes encoded clinical features, which are fused later in the network., which are fused later in the network. This setup enjoys the benefits of multi-modal information while keeping each branch specialised and well-structured.

Nair studied multimodal deep-learning models for cancer stage prediction using TCGA data [Nair \(2024\)](#). The study used transcriptome profiling, copy-number variation, and DNA methylation data to predict cancer stage across different cancer types. This work is close to DriftNet because it also studies cancer stage prediction using multimodal molecular information. However, DriftNet focuses on DNA methylation-derived features and clinical information, and it adds supervised contrastive learning before the final classifier. Therefore, a Nair-style multimodal neural network was used as a paper-inspired baseline to test whether simple multimodal neural fusion can compete with DriftNet.

2.8 Evaluation Practices and Stability Analyses

Reliable evaluation is a critical challenge in medical machine learning. Most analyses rely on a single train-test split, which may result in overly optimistic or pessimistic views dependent on the random seed and small changes in data [Kohavi et al. \(1995\)](#). Cross-validation, training with different seeds, and reporting metrics like Cohen’s kappa and per-class performance give a more realistic picture of model behaviour [Wong & Yeh \(2019\)](#) [McHugh \(2012\)](#). Some recent works stress the need to report variability across runs, rather than a single best result, when datasets are of moderate size.

In this study, we follow these best practices by adopting a fixed stratified split for the main experiment, running DriftNet with multiple random seeds, and using strict 5-fold cross-validation. We report overall accuracy, per-stage precision, recall, F1-score, and Cohen’s kappa, and provide the distribution of performances across the seeds. This stable evaluation framework helps to establish evidence that observed gains from contrastive learning and the DriftNet architecture are robust and not due to chance.

2.9 Positioning DriftNet

Taken together, the literature shows that DNA methylation is a rich signal for cancer biology, TCGA enables pan-cancer analyses, VAEs and contrastive learning can produce useful representations, and deep and ensemble models can support clinical prediction. However, no existing work brings all of these components together to solve the specific problem of pan-cancer AJCC Stage I–III classification using real patient data. Prior methylation-based models rarely target staging, VAEs alone do not directly separate cancer stages, contrastive learning has not been systematically applied to VAE-compressed methylation for pan-cancer staging, and multi-modal methods typically do not use contrastive, task-aware embeddings.

DriftNet is designed to address these gaps. It compresses more than 450,000 CpG measurements into a compact 150-dimensional latent space using a VAE, applies supervised triplet-loss contrastive learning to learn a 128-dimensional stage-aware representation, and fuses this representation with clinical features using a dual-path neural network trained with focal loss and strong regularisation.

To position DriftNet clearly, we compare it with both classical and paper-inspired deep-learning baselines. The classical baselines include logistic regression, random forest, gradient boosting, and k-nearest neighbours. The paper-inspired baselines include a Nair-style multimodal neural network, a DBN-ELM-BP-inspired deep model, a CMLV-AE + ANN baseline, and an OmiVAE-style VAE classifier [Nair \(2024\)](#); [Amanzholova & Coşkun \(2024\)](#); [Munquad et al. \(2025\)](#); [Zhang et al. \(2019\)](#). These models were selected because they represent important related ideas, including multimodal fusion, hybrid deep stage prediction, autoencoder-based latent feature learning, and VAE-based omics classification. Unlike these baselines, DriftNet uses supervised contrastive learning to reshape the representation before final Stage I/II/III classification.

Chapter 3

Methodology

This section describes the methodological framework used to develop DriftNet, detailing how high-dimensional methylation profiles and routine clinical variables were processed and integrated to produce reproducible Stage I–III predictions. It describes the compression of high-dimensional methylation data, the encoding of simple clinical features, how contrastive learning forms stage-aware embeddings, and how these pieces are combined into a two-branch neural network whose predictions of Stage I, II, and III are transparent and reproducible. The proposed model is shown in Figure 3.1.

Complete workflow of DriftNet, a deep-learning model for predicting cancer stage (I/II/III) using TCGA methylation-derived and clinical data. The pipeline includes (1) integration and harmonisation of data, (2) compression of high-dimensional DNA methylation into autoencoder-derived latent features and encoding of clinical variables, (3) contrastive learning using a Siamese triplet network to learn stage-aware embeddings, (4) a dual-branch neural network that combines contrastive embeddings with clinical information, and (5) model evaluation using repeated random-seed training, strict 5-fold cross-validation, and baseline comparisons.

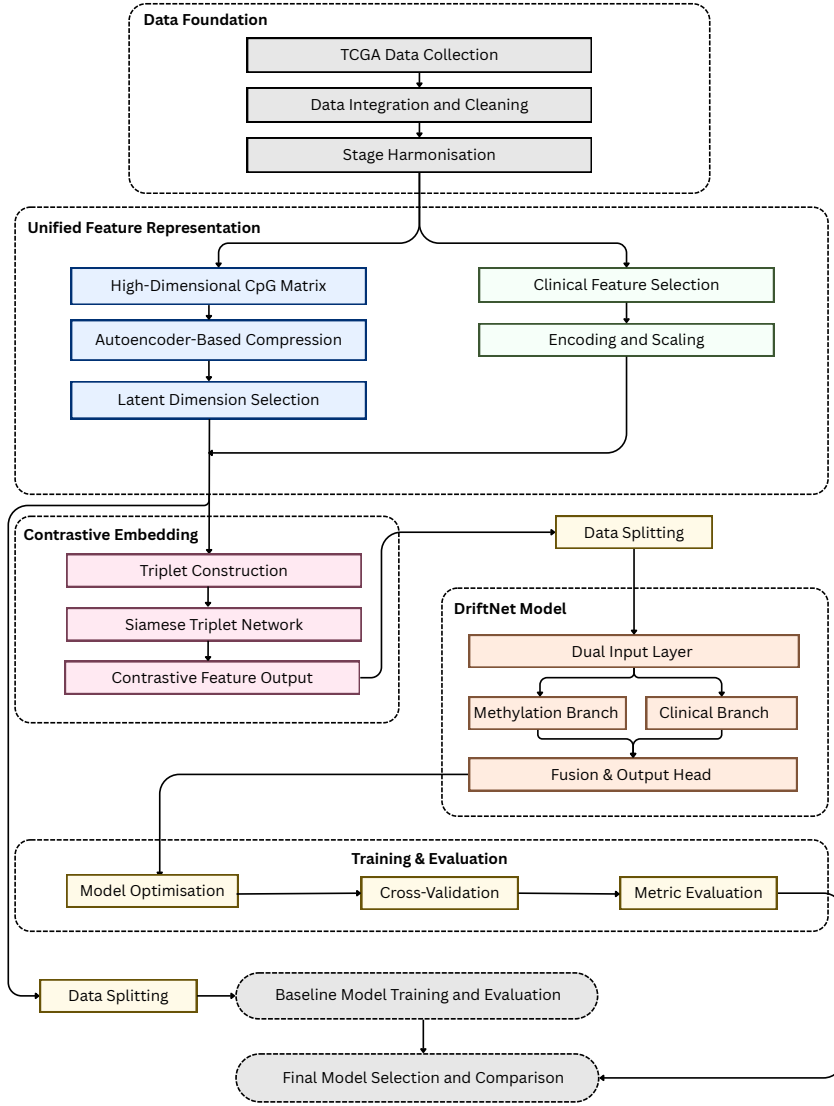


Figure 3.1: Overall DriftNet methodology workflow.

3.1 Autoencoder-Based Methylation Representation

The raw methylation data contains approximately 450,000 CpG sites per sample, which is too high-dimensional for direct model training. To compress this data into a more practical representation, we trained an autoencoder on the methylation matrix. The encoder learned a compact latent representation of each methylation

profile, while the decoder attempted to reconstruct the original CpG values from this compressed representation.

Several latent-space sizes were tested, including 2, 100, 150, 200, 250, 300, 350, 400, 450, and 500 dimensions. For each latent size, reconstruction loss was measured to assess how well the autoencoder could rebuild the original methylation profile. Among the tested sizes, the 150-dimensional latent space was selected because it provided a compact representation while maintaining strong reconstruction performance. Larger latent spaces increased model complexity but did not provide enough additional benefit for the final downstream pipeline.

The final autoencoder architecture used fully connected encoder layers to reduce the input methylation profile into a 150-dimensional latent vector. The decoder then reconstructed the input from this latent vector, and the model was trained using mean squared error loss. Because the methylation data was stored in multiple chunk files, the same autoencoder was trained continuously across all chunks rather than training separate models for each chunk. A global min-max scaling strategy was used so that all chunks followed the same scaling range. After training, the encoder was used to extract 150 latent features for every sample, and all chunk-level latent outputs were combined into one final latent-feature table for downstream modelling.

3.2 Clinical Feature Encoding

In addition to methylation, we employed 13 straightforward clinical variables commonly measured by physicians: sex, race, ethnicity, age at diagnosis, disease type, primary diagnosis, tissue or organ of origin, tumour morphology, tissue type, sample type, prior malignancy, prior treatment, and vital status.

Categorical variables (such as gender or race) were transformed into numbers via label encoding. Continuous variables (such as age) remained as numbers. To be used with the classic machine-learning baselines, we combined all 150 methylation features with these 13 clinical features into one long vector per patient and standardized all values to have zero mean and unit variance. For the classical and paper-based baseline models, the 150 autoencoder-derived methylation latent features and 13 encoded clinical features were combined into one pre-contrastive

feature vector. For DriftNet, the contrastive encoder first transformed the preprocessed features into a 128-dimensional stage-aware embedding, while the 13 clinical features were scaled separately using a standard scaler fitted only on the training set.

3.3 Triplet-Based Contrastive Feature Learning

Although the autoencoder produces a compact 150-dimensional representation of methylation features, the resulting space does not fully resolve the separation between Stage I, II, and III samples. There is considerable overlap remaining between Stage II and Stage III. In order to make the stages more separable, we then employed contrastive learning with triplet loss. The basic idea is straightforward: it is to place samples from the same stage closer together and push samples from different stages farther apart.

We implemented a Siamese encoder network that accepts a 163-dimensional pre-contrastive feature vector, consisting of 150 autoencoder-derived methylation features and 13 encoded clinical features, and returns a 128-dimensional stage-aware embedding. During training, we input three samples at a time into the encoder: an anchor, a positive from the same stage, and a negative from a different stage. Triplet loss pushes the anchor to be closer to the positive relative to the negative by a fixed margin, helping the network learn stage-aware representations.

3.4 DriftNet Stage Classification Model

DriftNet is a dual-branch neural network that combines the learned contrastive embedding with clinical features for final Stage I/II/III prediction. The first branch receives the 128-dimensional stage-aware contrastive embedding. A small Gaussian noise layer is applied to improve robustness, followed by a dense layer with 64 neurons, ReLU activation, L2 regularisation, and dropout.

The second branch receives the 13 encoded clinical features. This branch uses a dense layer with 32 neurons, ReLU activation, L2 regularisation, and dropout. The outputs of the embedding branch and the clinical branch are then concatenated and passed into a fusion layer with 64 neurons, ReLU activation, L2 regularisation, and

Algorithm 1 Triplet-Based Contrastive Feature Learning (DriftNet-Emb)

Input:

Pre-contrastive feature matrix $X \in \mathbb{R}^{N \times 163}$, where $X = [Z, C]$
Autoencoder methylation latent matrix $Z \in \mathbb{R}^{N \times 150}$
Encoded clinical feature matrix $C \in \mathbb{R}^{N \times 13}$
Stage labels $y \in \{\text{Stage I, Stage II, Stage III}\}$
Number of triplets T ; margin α

- 1: Standardise X using the scaler fitted for the contrastive-learning stage
- 2: Initialise encoder $f_\theta : \mathbb{R}^{163} \rightarrow \mathbb{R}^{128}$
- 3: Generate triplet set $\mathcal{T} = \{(x_a, x_p, x_n)\}_{t=1}^T$ using labels y
- 4: **repeat**
- 5: Shuffle $\mathcal{T}$ and form mini-batches
- 6: **for** each batch B **do**
- 7: $e_a \leftarrow f_\theta(x_a), e_p \leftarrow f_\theta(x_p), e_n \leftarrow f_\theta(x_n)$
- 8: Normalise e_a, e_p, e_n using L_2 normalisation
- 9:
$$L \leftarrow \frac{1}{|B|} \sum_{(x_a, x_p, x_n) \in B} \max(0, \|e_a - e_p\|^2 - \|e_a - e_n\|^2 + \alpha)$$
- 10: Update θ with Adam to minimise L
- 11: **end for**
- 12: **until** validation loss stops improving
- 13: Compute $E_i \leftarrow f_\theta(X_i)$ for all $i = 1, \dots, N$

Output:

Trained encoder f_θ ; stage-aware embedding matrix $E \in \mathbb{R}^{N \times 128}$

dropout. Finally, a dense softmax layer with three outputs predicts the probabilities for Stage I, Stage II, and Stage III.

The model was trained using sparse categorical cross-entropy. To reduce the effect of class imbalance, class weights were computed from the training set and then boosted for Stage II and Stage III, with the strongest boost applied to Stage III. This was done because Stage III had fewer samples and was clinically important to detect.

3.5 Baseline and Paper-Based Validation Models

To evaluate DriftNet fairly, we compared it with both classical machine-learning baselines and paper-inspired deep-learning baselines. The classical baselines were

logistic regression, random forest, gradient boosting, and k-nearest neighbours. These models were trained using the 150 autoencoder-derived methylation features and 13 clinical features before contrastive learning.

Four paper-inspired deep-learning baselines were also implemented. First, the Nair-style multimodal neural network used two branches: one branch for methylation features and another branch for clinical features. The two branches were then joined for final stage prediction. Second, the DBN-ELM-BP-inspired baseline used stacked dense layers to simulate deep feature extraction, followed by an ELM-like hidden layer and backpropagation-based training. Third, the CMLV-AE + ANN baseline used an autoencoder to learn compact latent features and then used an ANN classifier for Stage I/II/III prediction. Fourth, the OmiVAE-style classifier used a VAE-like representation learning structure with a classification head.

These baselines were trained on the 163-dimensional pre-contrastive feature vector, while DriftNet used the learned 128-dimensional contrastive embedding together with the scaled clinical features. This design was chosen because contrastive learning is the main contribution of DriftNet. If the baselines used DriftNet’s contrastive embeddings, the comparison would not be fully independent.

Algorithm 2 DriftNet Pan-Cancer Stage Classification Model

Input:

- Contrastive embeddings $E \in \mathbb{R}^{N \times 128}$
 - Clinical features $C \in \mathbb{R}^{N \times 13}$
 - Stage labels y and train/validation/test split
- 1: Split E , C , and y into training, validation, and test sets
 - 2: Standardise clinical features using the training set statistics
 - 3: Build embedding branch on E with Gaussian noise, dense layer, L2 regularisation, and dropout
 - 4: Build clinical branch on C with dense layer, L2 regularisation, and dropout
 - 5: Concatenate both branch outputs
 - 6: Add fusion dense layer with L2 regularisation and dropout
 - 7: Add 3-way softmax output layer for Stage I, Stage II, and Stage III
 - 8: Compute class weights from y_{train}
 - 9: Apply additional weight boosts for Stage II and Stage III
 - 10: Define sparse categorical cross-entropy loss with class weights
 - 11: Initialise Adam optimiser
 - 12: **repeat**
 - 13: Sample mini-batch (E_b, C_b, y_b) from the training set
 - 14: $\hat{p}_b \leftarrow g_\phi(E_b, C_b)$
 - 15: $L \leftarrow \text{CrossEntropy}(\hat{p}_b, y_b)$ with class weights
 - 16: Update ϕ with Adam to minimise L
 - 17: Evaluate validation accuracy and update early stopping / learning-rate scheduler
 - 18: **until** validation accuracy does not improve
 - 19: Select the validation-best parameters ϕ^*
 - 20: Evaluate g_{ϕ^*} on the held-out test set

Output:

- Trained DriftNet $g_{\phi^*}(E, C)$ for Stage I/II/III prediction
-

Chapter 4

Experiments and Result Analysis

This section describes how DriftNet was evaluated and summarises the resulting performance. First, we describe the TCGA cohorts used in the study and the procedures applied to clean, merge, and filter the methylation and clinical data to obtain a reliable Stage I–III dataset. We then detail the training, validation, and test splits, along with the standard settings that were used to ensure that other researchers can repeat the work. Finally, it walks through the main results step by step: how the autoencoder compresses methylation data, how contrastive learning helps in separating stages, how well DriftNet predicts stage compared with baseline models, and the stability and interpretability of these results across many runs.

4.1 Dataset and Pre-processing

The present study uses patient data from The Cancer Genome Atlas. For every patient in TCGA, there are two types of information:

- DNA methylation measurements (many CpG sites)
- Basic clinical data such as age, cancer type, and stage.

We combined all files so each row in the resulting table is one patient with both methylation and clinical features. Several filtering steps were performed on the data to maintain quality and ensure results can be reproduced with reliable accuracy. Starting with 7,465 raw samples from TCGA methylation arrays, after autoencoder-based methylation compression and initial quality checks, we retained

Table 4.1: Final cohort distribution by diagnosis and stage.

Primary diagnosis	Stage I	Stage II	Stage III	Total	Percent
Breast, NOS	84	295	132	511	38.05
Thyroid gland	208	30	71	309	23.01
Kidney, NOS	171	37	76	284	21.15
Upper lobe, lung	147	64	28	239	17.80
Total	610	426	307	1343	100.00

7,404 samples that had complete methylation profiles, removing 61 samples whose CpG measurements were missing or incomplete.

Then, we filtered the data by clinical stage to include only Stage I, Stage II, and Stage III cases, as these stages have the most relevance for treatment decisions. After filtering, there were 2,624 samples with valid stage labels, while 4,780 samples lacked stage information, were labelled as Stage 0 or Stage IV, or had ambiguous staging annotations. Stage 0 (in situ) and Stage IV (metastatic) cases were intentionally excluded because they represent distinct biological states that require separate analytical frameworks. Samples containing missing key clinical values or latent methylation features were removed, leaving a clean set ready for modelling, shown in Table 4.1. Only primary diagnosis categories with ≥ 238 samples were retained, yielding 4 diagnoses and 1,343 patients for analysis.

We imposed a minimum sample size threshold for each category of primary diagnosis to ensure statistical stability and prevent overfitting from small sample groups. Among 66 unique primary diagnosis categories in our dataset, we excluded 62 smaller primary diagnosis categories; thus, 1,281 samples were removed. This filtering is an important step, as the cancer types with very small representation may introduce noise and provide unreliable performance estimates, compromising the generalizability of this model. The final analysis cohort consisted of 1,343 patients across 4 well-represented primary diagnoses, ensuring that each diagnostic category meaningfully contributes to the model training and evaluation. The four retained primary diagnosis groups were Breast, NOS; Thyroid gland; Kidney, NOS; and Upper lobe, lung. These groups were retained because each had enough samples to support stable training and evaluation.

In addition, strings like “-”, “unknown”, and “not reported” were converted to proper missing values. Columns that were mostly missing were removed. Of

Table 4.2: Input features used for modelling

Feature group	Number of features
Latent methylation features	150
Clinical features	13
Total per patient	163

the remaining clinical fields, 13 simple but important variables were retained; for example, gender, age, primary diagnosis, tissue of origin, prior malignancy, prior treatment, and vital status. Only those patients with no missing values in these 13 variables were retained, and the 150 latent methylation features were kept for the experiments. (Table 4.2)

Categorical clinical variables were encoded as integers; numerical ones such as age were kept as continuous values. All features were standardised (zero mean, unit variance) before training.

4.2 Experimental Setting

For the supervised experiments, the 1,343 labelled patients were split into three non-overlapping groups by stratified sampling, so stages I, II, and III remained roughly the same across each group. The final allocation was as follows:

- 70% (940 patients) for training.
- 15% (201 patients) for validation and model tuning.
- 15% (202 patients) for independent testing.

The contrastive encoder was trained to learn stage-aware representations for DriftNet. For fair evaluation, the classical machine-learning baselines and paper-inspired deep-learning baselines were trained using the pre-contrastive feature set, consisting of 150 autoencoder-derived methylation features and 13 clinical features. DriftNet was the only model trained using the 128-dimensional supervised contrastive stage-aware embeddings together with the 13 clinical features. All models used the same train-validation-test split to ensure a fair comparison. This design ensures that the baseline models do not receive the benefit of DriftNet’s contrastive learning module.

Training was performed using mini-batches, the Adam optimizer with a fixed learning rate, and early stopping based on validation accuracy to prevent overfitting. Every experiment was run several times with different random seeds, and the results were presented by showing the distribution of test accuracies across the runs for stability, not just the best single number.

4.3 Evaluation Metrics

This study uses simple and standard metrics to allow the readers to clearly assess how well DriftNet and the baseline models perform for each stage of cancer. All metrics were derived from the confusion matrix of the held-out test set and were averaged across multiple runs to account for variability.

- Overall accuracy: The proportion of test patients who receive the correct stage label. Easy to understand and allows us to quickly compare models.
- Precision: For one stage (say Stage III), precision asks: “Among the patients the model calls Stage III, how many are actually Stage III?” High precision means few false alarms for that stage.
- Recall / Sensitivity (per stage): For a given stage, recall quantifies the proportion of true cases correctly identified by the model. High recall means the model does not miss many cases in that stage, which matters for both early and late stages.
- F1-score (by stage): This shows the harmonic mean of precision and recall for each stage. It is only high if both precision and recall are good. Therefore, balancing false positives and false negatives.
- Macro-average F1: It is the simple average of three stage-wise F1 scores. It gives equal importance to all stages, irrespective of the number of samples in each stage, and reflects how well a model performs across various stages in a balanced manner.
- Cohen’s kappa: While similar to accuracy, this measures the amount by which the model’s predictions agree with the true labels over and above agreement that could occur by chance. Values closer to one mean strong agreement;

values around 0.5–0.6 are usually seen as “moderate to good” in medical work.

- **Confusion matrix:** A 3×3 table displaying, for each true stage, the number of patients predicted as Stage I, II, or III. It helps identify common mistakes, especially how often Stage II is mixed up with Stage III and how rarely Stage I is mistaken for Stage III.

4.4 Analysis of Experimental Results

This section provides a description of what the experiments show about the quality of our representations, how accurate DriftNet is, and how reliable the results are. First, we investigate how the autoencoder compresses the very large methylation matrix into a 150-dimensional space and how we choose this latent size. Then, we will demonstrate that contrastive learning reshapes this space such that the overall stage-aware structure becomes clearer, both visually and according to clustering metrics. Next, we will present the main staging results for DriftNet, including per-stage precision, recall, F1-scores, and overall accuracy, and check how stable these numbers are across many random seeds and cross-validation folds. We further compare DriftNet with several strong baseline models, which were trained on exactly the same features, in order to see how much benefit the proposed architecture and contrastive embeddings bring. Finally, we analyze confusion patterns and representative failure cases to understand where misclassifications occur, why most confusion arises between adjacent stages, and how these insights may guide future methodological improvements.

4.4.1 Autoencoder Compression and Latent Dimension Selection

Each sample in the raw methylation data contains approximately 450,000 CpG sites, which is too large to directly feed into a regular classifier. To reduce dimensionality while retaining important biological signal, we trained an autoencoder to compress the CpG profiles into a smaller latent representation. The encoder maps the high-dimensional methylation profile into a low-dimensional latent vector, while the decoder reconstructs the original CpG values from this compressed code. The

model was trained by minimising reconstruction error, which indicates how well the latent space preserves the original methylation information.

Several latent sizes were evaluated: 100, 150, 200, 250, 300, 350, 400, 450, and 500. For each latent size, the average reconstruction loss was recorded. As shown in Figure 4.1, the 150-dimensional latent representation provided the best balance between compactness and reconstruction quality, and it was therefore used as the standard methylation representation in later experiments.

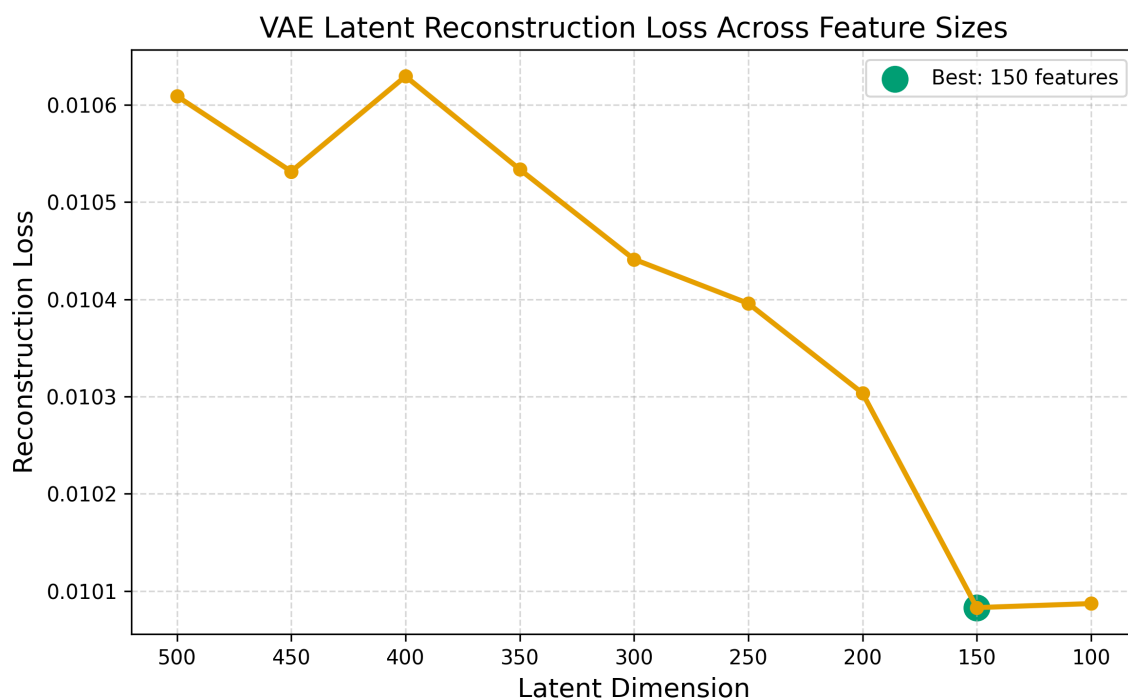


Figure 4.1: Autoencoder reconstruction loss across latent dimensions.

4.4.2 Contrastive Learning Improves Stage Separability

Although the 150-dimensional autoencoder features provide a compact representation of the methylation profiles, they do not produce clear separation between Stage I, II, and III samples. In fact, as we project the autoencoder latent space into a two-dimensional scatter plot with t-SNE, Stage II and III samples exhibit a high degree of overlap, as shown in the Figure 4.2 left panel. Hence, the differences between the two stages are hard for any classifier to distinguish.

To improve separability, we employed supervised triplet-loss contrastive learning. The intuition is simple: train a neural network, the contrastive encoder, to reshape the 163-dimensional pre-contrastive feature space into a new 128-dimensional stage-aware space where samples from the same stage are pulled closer and samples from different stages are pushed apart. We generated triplets of samples: an anchor, a positive and a negative. The loss function enforces that the anchor-positive distance is smaller than the anchor-negative distance by a fixed margin. We generated 20,000 triplets from 1,343 patient samples to ensure adequate representation of all three stages and trained the Siamese encoder for up to 100 epochs with early stopping based on validation loss.

After training, the encoder was used to obtain 128-dimensional contrastive embeddings for all samples, and the quality of the representations obtained was examined through several clustering metrics. Compared with the original pre-contrastive feature space, the contrastive embeddings demonstrated:

- Silhouette score improved from -0.0185 in the original feature space to 0.1461 in the contrastive space, indicating better overall stage grouping.
- Davies–Bouldin index decreased from 5.5563 to 2.4061, where lower values indicate more compact and better separated clusters.
- Calinski–Harabasz score increased from 68.4 to 281.9, showing that the contrastive embeddings improved cluster separation.
- LDA classification score across all three stages improved from 0.6739 to 0.7453, indicating better overall class discrimination after contrastive learning.
- Stage I–II centroid cosine similarity decreased from 0.9965 to 0.7996, showing better separation between these two stages. However, Stage I–III similarity increased from 0.0514 to 0.7281, and Stage II–III similarity increased from 0.0328 to 0.3684. Therefore, the centroid similarity results are mixed. The strongest evidence for improved representation quality comes from the Silhouette Score, Davies–Bouldin Index, Calinski–Harabasz Score, and LDA score.

Overall, the contrastive learning stage improved the global structure of the feature space, as shown by better Silhouette, Davies–Bouldin, Calinski–Harabasz, and LDA scores. However, the cosine-similarity results show that adjacent cancer stages

still remain difficult to separate perfectly, especially Stage II and Stage III. This is consistent with the final classification and error-analysis results.

As illustrated in Figure 4.2, the t-SNE projection of the original 150-dimensional autoencoder features shows substantial overlap among the stages, especially around Stage II and Stage III. After applying contrastive learning, the 128-dimensional embeddings show clearer stage-wise structure and improved global cluster separation. However, Stage II and Stage III still remain difficult to separate completely, which is consistent with the classification and error-analysis results.

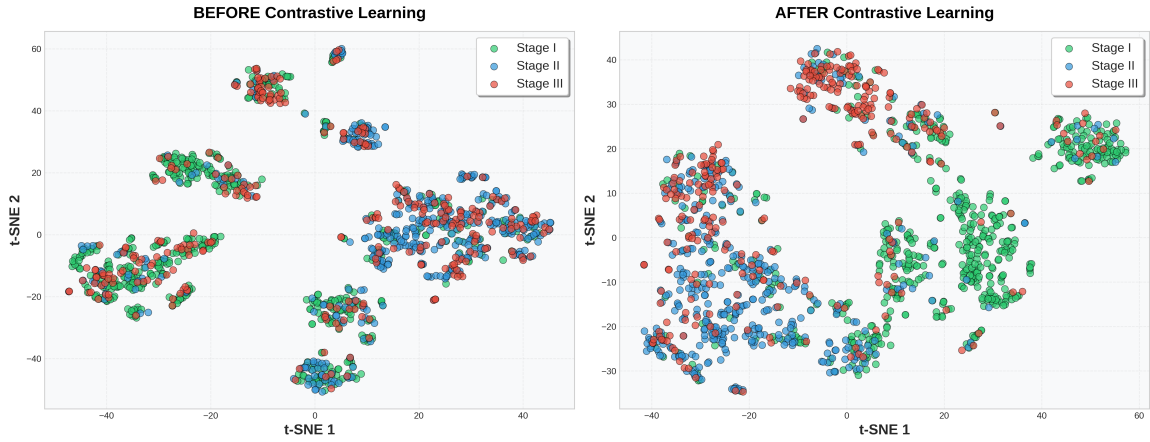


Figure 4.2: A side-by-side t-SNE plot.

4.4.3 DriftNet Stage Classification Performance

DriftNet is a dual-branch neural network that uses the 128-dimensional contrastive stage-aware embedding and 13 clinical features to predict cancer stage. The embedding branch applies Gaussian noise, a dense layer with 64 neurons, L2 regularisation, and dropout. The clinical branch processes the 13 clinical variables using a dense layer with 32 neurons, L2 regularisation, and dropout. The outputs of both branches are concatenated and passed through a fusion layer with 64 neurons before the final softmax layer predicts Stage I, Stage II, or Stage III.

The model was trained using the Adam optimiser and sparse categorical cross-entropy. Class weights were computed from the training set and then boosted for Stage II and Stage III, with the strongest boost applied to Stage III. Early stopping monitored validation accuracy, and the best model was selected based on validation

Table 4.3: DriftNet performance on the held-out test set.

Stage	Precision	Recall	F1-Score	Accuracy
Stage I	0.8539	0.8261	0.8398	82.61% (76/92)
Stage II	0.6600	0.5156	0.5789	51.56% (33/64)
Stage III	0.4444	0.6087	0.5138	60.87% (28/46)
Macro Avg	0.6528	0.6501	0.6442	
Weighted Avg	0.6992	0.6782	0.6829	67.82%
Overall Accuracy	67.82%			

performance. The final fixed split used 940 training samples, 201 validation samples, and 202 held-out test samples. The main performance metrics are presented in Table 4.3.

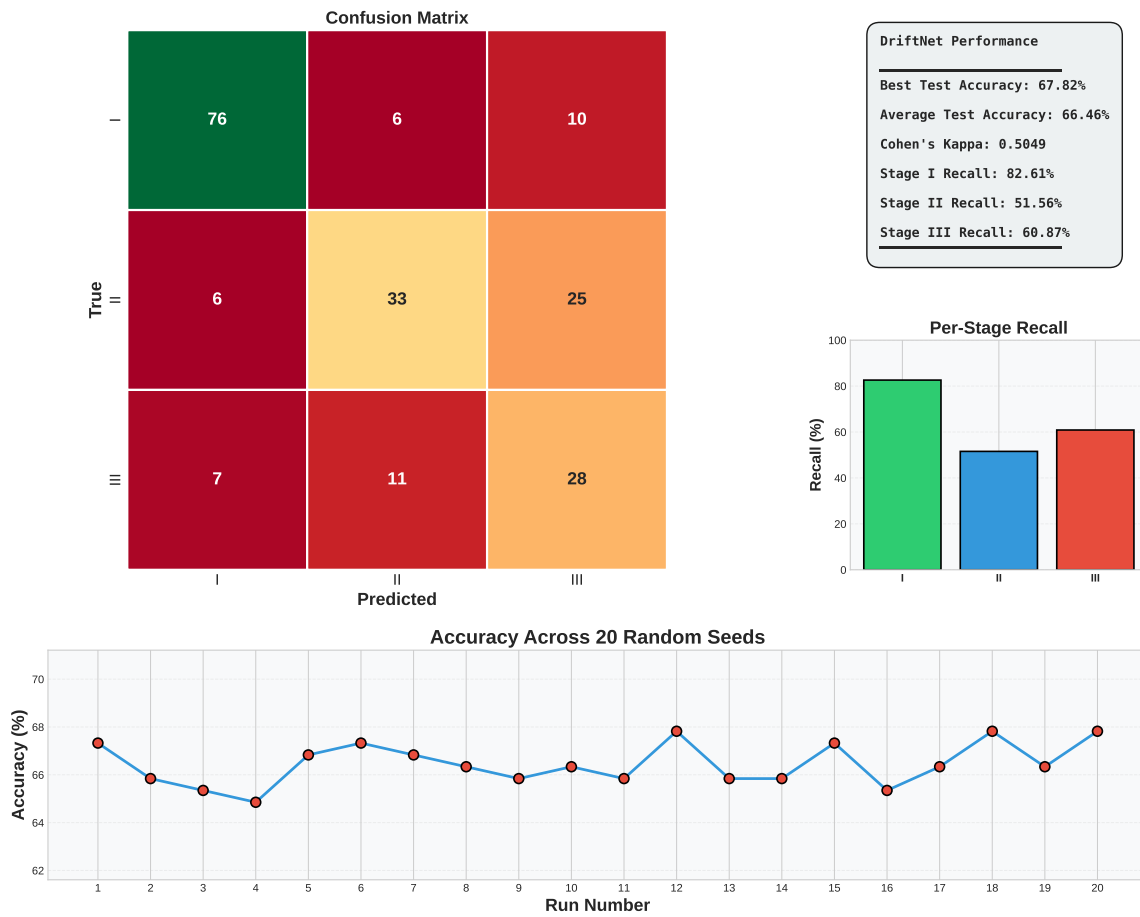


Figure 4.3: DriftNet Performance Dashboard.

(Figure 4.3) provides a comprehensive view of DriftNet model results on the test set. This dashboard presents the confusion matrix, a summary of important metrics including accuracy, kappa and recall at each stage, a bar plot of recall at every stage, and a line plot presenting test accuracy over 20 runs with random seeds. This figure illustrates both the high performance and the stability of the model across various test scenarios. Together, these results show that DriftNet achieved stable performance across repeated seed runs and provides useful stage-wise prediction behaviour on the held-out test set.

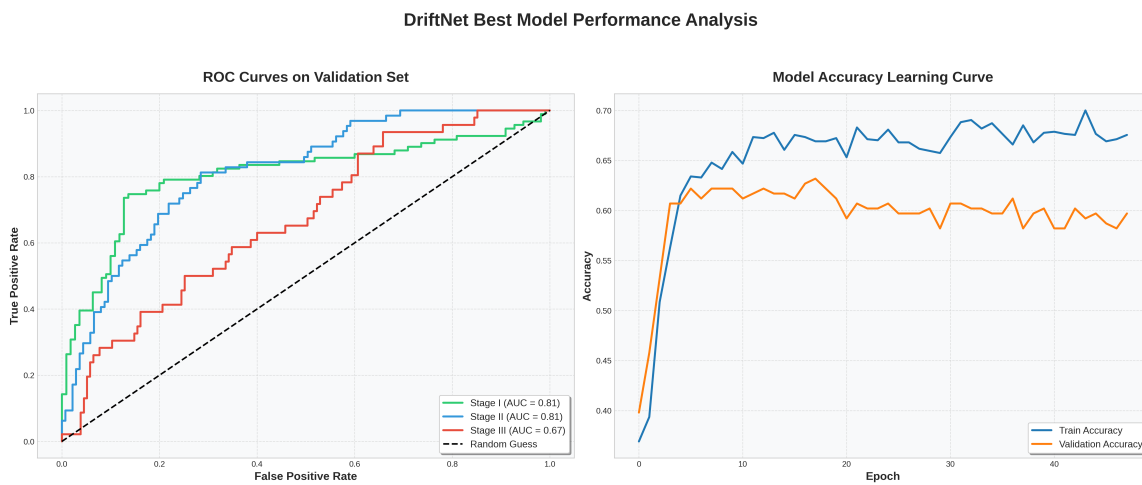


Figure 4.4: ROC curves and learning curve of the best DriftNet model.

Figure 4.4 shows the ROC curves for the three stages and the training-validation accuracy curve for the best DriftNet model. Stage I and Stage II achieved AUC values of 0.81, while Stage III achieved a lower AUC of 0.67. This indicates that Stage III remained harder to distinguish from the other stages. The learning curve shows that training accuracy increased and stayed above validation accuracy, while validation accuracy remained around 0.60–0.63. This suggests that the model learned useful patterns, but the gap between training and validation performance also indicates some remaining generalisation difficulty.

4.4.4 Cross-Validation Analysis

We conducted a 5-fold stratified cross-validation on the entire group of 1,343 samples. In each fold, 80% of the data was used for training and 20% for testing, with

Table 4.4: Five-fold stratified cross-validation results.

Stage	Precision	Recall	F1-Score	Accuracy
Stage I	0.8435	0.7246	0.7795	72.46% (442/610)
Stage II	0.6950	0.5188	0.5941	51.88% (221/426)
Stage III	0.4371	0.7134	0.5421	71.34% (219/307)
Macro Avg	0.6585	0.6522	0.6386	—
Weighted Avg	0.7035	0.6567	0.6664	—
Overall Accuracy				65.67% $\pm$ 2.62%

splits kept to have similar stage ratios. A new DriftNet model was trained from scratch for every fold. The five-fold stratified cross-validation results are presented in Table 4.4.

Figure 4.5 provides a visual summary of the cross-validation performance:

- Left panel: A violin or box plot of accuracy across the 5 folds.
- Middle panel: An aggregated confusion matrix for all 5 folds.
- Right panel: A grouped bar chart of the per-stage recall for each fold and the overall mean; this is to show stable per-class performance.

Together, these results demonstrate that DriftNet achieves stable performance across repeated resampling of the data.

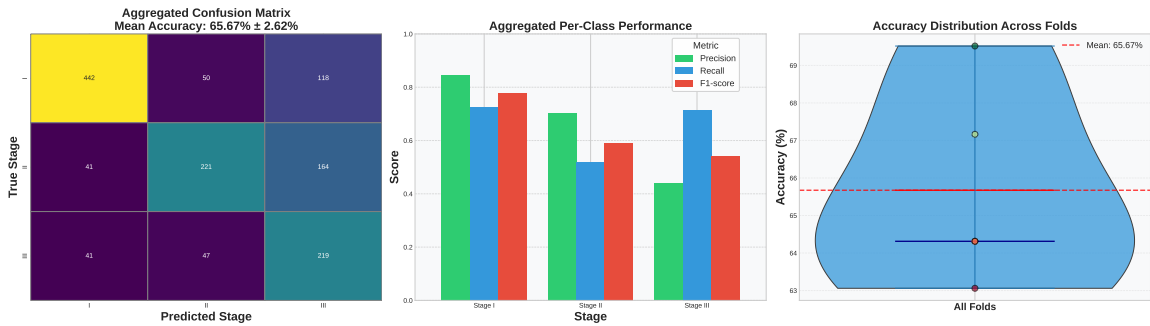


Figure 4.5: Five-fold cross-validation performance summary.

4.4.5 Comparison with Classical and Paper-Based Baseline Models

To assess the added value of DriftNet, we compared it with two groups of baseline models. The first group consisted of classical machine-learning models, including

Table 4.5: Final comparison of DriftNet with comparison models.

Rank	Model	Model Group	Accuracy	Kappa	Macro F1	Weighted F1	Stage III F1
1	DriftNet	Proposed	67.82%	0.5049	0.6442	0.6829	0.5138
2	CMLV-AE + ANN	Paper-Based Deep	64.85%	0.4517	0.6040	0.6483	0.4045
3	Gradient Boosting	Regular ML	65.84%	0.4564	0.5996	0.6467	0.3797
4	Logistic Regression	Regular ML	62.87%	0.4200	0.5816	0.6282	0.3778
5	DBN-ELM-BP-inspired	Paper-Based Deep	61.39%	0.4075	0.5810	0.6213	0.3958
6	Nair-style Multimodal NN	Paper-Based Deep	62.38%	0.4100	0.5714	0.6182	0.3373
7	Random Forest	Regular ML	63.37%	0.4032	0.5398	0.5979	0.2222
8	Omiautoencoder-style autoencoder Classifier	Paper-Based Deep	61.39%	0.3585	0.4686	0.5447	0.0400
8	KNN	Regular ML	60.40%	0.3521	0.5336	0.5805	0.3158

Logistic Regression, Random Forest, Gradient Boosting, and k-Nearest Neighbours. The second group included paper-inspired deep-learning baselines: a Nair-style Multimodal Neural Network, a DBN-ELM-BP-inspired model, a CMLV-AE + ANN model, and an Omiautoencoder-style autoencoder classifier.

All baseline models were trained using the same pre-contrastive feature representation, which consisted of 150 autoencoder-derived methylation features combined with 13 clinical features. DriftNet, in contrast, used the 128-dimensional supervised contrastive stage-aware embeddings together with the same 13 clinical features. This evaluation design ensures that the baseline models are not given access to the proposed contrastive representation. Therefore, the comparison reflects whether the contrastive learning stage and the dual-branch DriftNet architecture provide additional predictive value beyond standard autoencoder-based features.

As shown in Table 4.5, DriftNet achieved the strongest overall performance among all evaluated models. It obtained the highest test accuracy of 67.82%, the highest Cohen’s kappa of 0.5049, the highest macro F1-score of 0.6442, and the highest Stage III F1-score of 0.5138. These results indicate that DriftNet performed well not only in overall classification accuracy, but also in balanced stage-wise performance.

Among the classical machine-learning baselines, Gradient Boosting achieved the highest accuracy, with 65.84%. However, its Stage III F1-score was 0.3797, which was considerably lower than DriftNet’s Stage III F1-score. Among the paper-inspired deep-learning baselines, CMLV-AE + ANN achieved the best performance, with 64.85% accuracy and a Stage III F1-score of 0.4045. Although these baselines performed competitively, neither matched DriftNet’s overall ranking or its performance on the Stage III class.

This comparison suggests that the supervised contrastive learning stage contributes meaningfully to the final classification performance. While the baseline models used the 150-dimensional autoencoder latent features directly, DriftNet first transformed these features into a 128-dimensional stage-aware embedding before classification. This additional representation-learning step appears to improve the separation of cancer stages, particularly for Stage III, which remains the most challenging class in this study.

4.4.6 Impact of Dimensionality Reduction: Comparing Full dimensional and Autoencoder-based Latent Space

One important question in this study is whether compressing methylation data helps or harms prediction performance. The original methylation data contains 486,362 CpG measurements per patient, which is very high-dimensional compared with the number of available patients. Training directly on such a large feature space can increase computational cost and may also cause the model to learn noise instead of useful biological patterns. For this reason, we used an autoencoder to reduce the methylation profile into a compact 150-dimensional latent representation.

To examine the effect of this dimensionality reduction, we compared two feature settings. The first setting used the original full-dimensional methylation features together with 13 clinical features. The second setting used the 150 autoencoder-derived methylation features together with the same 13 clinical features. This comparison helps show whether the compressed autoencoder representation provides a more useful input space for stage classification.

Table 4.6 summarizes the results. The full-dimensional feature setting achieved 56.13% test accuracy, while the autoencoder-based representation achieved 67.82% test accuracy. This shows an improvement of 11.69 percentage points after using autoencoder-based dimensionality reduction. The difference was also visible in the stage-wise recall values. The full-dimensional setting achieved 81.15% recall for Stage I, 55.29% for Stage II, and only 8.06% for Stage III. In contrast, the autoencoder-based setting achieved 82.61% recall for Stage I, 51.56% for Stage II, and 60.87% for Stage III. The most important improvement was observed for Stage III, where recall increased from 8.06% to 60.87%.

These results indicate that the autoencoder representation provides a more useful and stable feature space than the original full-dimensional CpG matrix.

Table 4.6: Performance comparison of full-dimensional and autoencoder features.

Metric	Full-dimensional Features	autoencoder-based Features
Test Accuracy	56.13%	67.82%
Stage I Recall	0.8115	0.8261
Stage II Recall	0.5529	0.5156
Stage III Recall	0.0806	0.6087
Macro F1	0.4561	0.6442

Although Stage II recall was slightly lower in the autoencoder-based setting, the large improvement in Stage III recall shows that the compressed representation helped the model detect advanced-stage cases much better. This is important because Stage III is one of the most difficult classes in this study.

The full-dimensional methylation space contains far more features than patient samples, which makes the model highly vulnerable to overfitting. In contrast, the 150-dimensional autoencoder representation reduces noise and keeps the most meaningful structure of the methylation profile. Therefore, autoencoder compression not only reduces computational cost and memory usage but also improves the model’s ability to generalize to unseen patients.

4.4.7 Error Analysis and Misclassification Patterns

Error analysis helps us understand where DriftNet performs well and where it still makes mistakes. Cancer stages are not always easy to separate because adjacent stages can share similar biological patterns. In this study, Stage II and Stage III are especially difficult because their molecular differences may be small, and important clinical details such as tumour size, lymph-node involvement, and exact anatomical spread may not be fully captured by methylation and basic clinical variables alone.

Table 4.7 shows the transposed confusion matrix for the final DriftNet model on the held-out test set. Out of 92 Stage I patients, 76 were correctly classified. For Stage II, 33 out of 64 patients were correctly classified. For Stage III, 28 out of 46 patients were correctly classified. The model achieved the highest recall for Stage I at 82.61%, while Stage III recall was 60.87%. Stage II was the most difficult class, with a recall of 51.56%.

Table 4.7: DriftNet misclassification patterns on the test set.

Predicted Stage	True Stage I	True Stage II	True Stage III
Pred. Stage I	76	6	7
Pred. Stage II	6	33	11
Pred. Stage III	10	25	28
% Correct	82.61%	51.56%	60.87%

The main source of error was confusion between neighbouring stages, especially Stage II and Stage III. For example, 25 Stage II cases were predicted as Stage III, while 11 Stage III cases were predicted as Stage II. This pattern is expected because Stage II and Stage III often represent adjacent levels of disease progression. Their difference may depend on clinical information that is not fully represented in the current feature set, such as tumour size, lymph-node status, and detailed pathological measurements.

DriftNet also made some wider-stage errors. Ten Stage I cases were predicted as Stage III, and seven Stage III cases were predicted as Stage I. These errors are important because they may represent larger differences in disease severity. Therefore, DriftNet should not be interpreted as a replacement for clinical staging. Instead, it should be considered a research model and a possible decision-support tool that may assist clinicians only after further validation.

Overall, the error analysis shows that DriftNet performs best on Stage I and achieves the strongest Stage III F1-score among the compared models. However, Stage II remains the most difficult class. Future work should include additional clinical variables, such as tumour size, lymph-node involvement, and metastasis-related details, to improve separation between Stage II and Stage III. Another useful direction is to add uncertainty estimation, so the model can flag low-confidence cases for expert review instead of making overconfident predictions.

Chapter 5

Design and Development

This chapter presents the design and development of the DriftNet system. The purpose of the system is to provide a complete research prototype that can accept one methylation file and a set of clinical features, process them through the trained DriftNet pipeline, and return a predicted cancer stage. The system was developed as a web-based platform so that the trained model can be used in a clear, organized, and user-friendly way.

The main design problem in this project is not only building a prediction model, but also creating a practical system that can connect data input, preprocessing, feature generation, classification, and result visualization in one complete workflow. Since the model uses both DNA methylation information and clinical features, the system must manage multiple input types while keeping the process simple for the user. To solve this problem, DriftNet was developed using a modular design that includes a frontend user interface, a backend processing layer, and the trained machine-learning inference pipeline.

5.1 Design Overview

The overall design of DriftNet follows a modular architecture. The frontend is responsible for user interaction, the backend handles validation and communication, and the machine-learning subsystem performs feature transformation and final prediction. This separation was used because it makes the system easier to organize, test, debug, and extend in the future.

The web platform begins with a homepage section that introduces the system and explains its purpose. Figure 5.1 shows the top section of the webpage. This part

presents the project title, a short description of the system, and the key technical highlights of the model, including the number of CpG probes, latent feature size, contrastive embedding size, and number of clinical features.

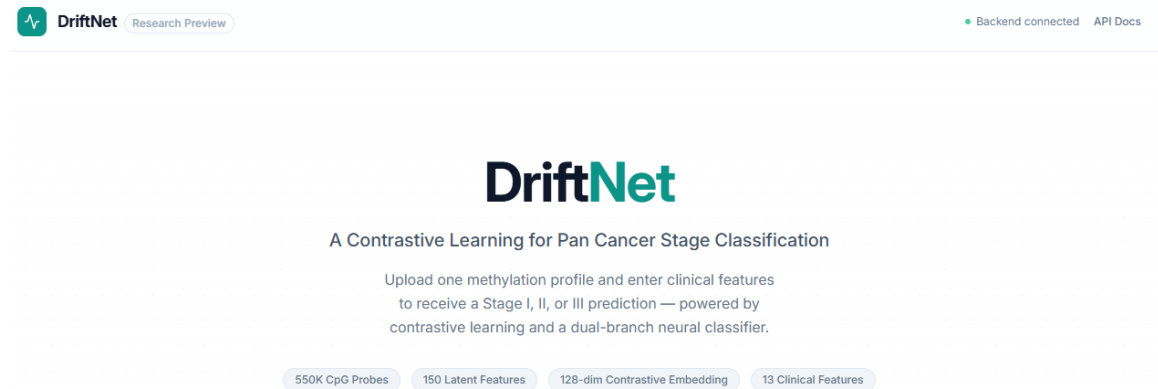


Figure 5.1: DriftNet homepage interface.

This homepage design helps users quickly understand what the system does before interacting with it. It also acts as the first communication layer between the technical model and the end user.

5.2 Web Interface Design

The DriftNet web interface was designed to guide the user step by step. Instead of exposing the model internals directly, the system presents the workflow in a clear sequence. This improves usability and reduces confusion.

5.2.1 Inference Workflow View

Figure 5.2 shows the four-step inference pipeline presented on the webpage. This design explains the complete flow of the system in a visual way. The four stages are: upload methylation file, enter clinical features, generate embeddings, and predict cancer stage.

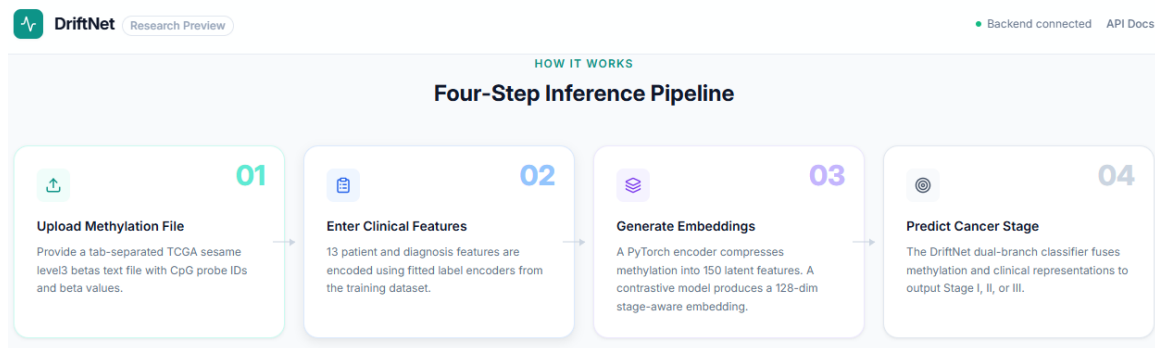


Figure 5.2: Four-step DriftNet inference workflow.

This workflow section is important because it increases transparency. Users can see that the result is not produced in one hidden step. Instead, the system processes the input through a structured pipeline that matches the actual inference logic of the trained model.

5.2.2 Main Prediction Page Layout

Figure 5.3 shows the main web view where the actual prediction process takes place. This page contains the upload panel, the clinical feature form, and the prediction controls. It serves as the central working area of the platform.

DriftNet Research Preview

Backend connected API Docs

RUN A PREDICTION

Upload & Classify

Provide a methylation file and complete the clinical form, then click Predict.

Methylation File

Tab-separated .txt — no header required

Drag & drop your file here

or [click to browse](#)

Accepted format: .txt
TCGA sesame level3 betas · EPIC array

Expected file format

```
cg00000029 0.125476
cg00000108 0.967176
cg00000807 NA
```

Clinical Features

All 13 fields are required 0/13

PATIENT DEMOGRAPHICS

GENDER •

Select...

RACE •

Select...

VITAL STATUS •

Select...

ETHNICITY •

Select...

SAMPLE INFORMATION

SAMPLE TYPE •

Select...

TISSUE TYPE •

Select...

DISEASE CLASSIFICATION

DISEASE TYPE •

Select...

DIAGNOSIS DETAILS

PRIMARY DIAGNOSIS •

Select...

TISSUE / ORGAN OF ORIGIN •

Select...

MORPHOLOGY CODE (ICD-O) • ⓘ

AGE AT DIAGNOSIS • ⓘ

[Predict Cancer Stage](#)

Figure 5.3: Main DriftNet prediction interface.

This layout was selected to keep all required interaction in one place. The left side is used for file upload, while the right side is used for clinical information. The prediction controls are placed below so that the full process remains easy to follow.

5.3 Major Subsystems in the Design

The complete system contains three major subsystems: the frontend interface subsystem, the backend processing subsystem, and the machine-learning inference subsystem.

5.3.1 Frontend Interface Subsystem

The frontend is the visible part of the system. Its purpose is to allow the user to interact with DriftNet without requiring knowledge of the internal model imple-

mentation. It includes the homepage, workflow explanation, input forms, upload areas, buttons, and the result display panel.

The frontend design focuses on clarity and simplicity. Important input fields are grouped into meaningful sections, such as patient demographics, sample information, disease classification, and diagnosis details. Many of the fields are implemented as dropdown menus to reduce typing mistakes and improve consistency.

5.3.2 Backend Processing Subsystem

The backend connects the user interface with the trained model. Its main role is to receive the methylation file and clinical values, validate the input, apply preprocessing, load the trained encoders and classifiers, run inference, and return the final prediction result to the frontend.

The backend also ensures that the same preprocessing steps used during training are reused during deployment. This is important because differences between training-time preprocessing and deployment-time preprocessing can reduce the reliability of predictions.

5.3.3 Machine-Learning Inference Subsystem

The machine-learning subsystem is the main analytical core of the system. It is responsible for transforming the raw inputs into the final cancer stage prediction. The inference pipeline includes three major operations:

- autoencoder-based methylation compression,
- contrastive embedding generation,
- final DriftNet stage classification.

First, the methylation file is converted into a numeric CpG beta-value vector. Then, the trained autoencoder encoder compresses this high-dimensional methylation input into a 150-dimensional latent representation. After that, the latent methylation features are combined with the encoded clinical variables to form the pre-contrastive feature vector. This vector is passed through the trained contrastive encoder to obtain a 128-dimensional stage-aware embedding. Finally, the DriftNet

classifier uses the embedding together with the clinical branch to predict Stage I, Stage II, or Stage III.

5.4 Input Design

The input part of the system must collect two types of data: a methylation file and 13 clinical features. Figure 5.4 shows the system after a methylation file has been uploaded and the clinical form has been completed.

The screenshot displays the DriftNet input interface. On the left, a file upload area shows a file named 'sample4826.txt' (12.6 MB) with a 'Remove file' option. Below it, an 'Expected file format' section shows a tab-separated text file structure with CpG probe IDs and beta values. On the right, a clinical form is filled out with the following data:

CLINICAL FEATURES	
GENDER: female	VITAL STATUS: Dead
RACE: white	ETHNICITY: not hispanic or latino
SAMPLE INFORMATION	
SAMPLE TYPE: Primary Tumor	TISSUE TYPE: Tumor
DISEASE CLASSIFICATION	
DISEASE TYPE: Adenomas and Adenocarcinomas	
DIAGNOSIS DETAILS	
PRIMARY DIAGNOSIS: Papillary adenocarcinoma, NOS	TISSUE / ORGAN OF ORIGIN: Thyroid gland
MORPHOLOGY CODE (ICD-O): 8260/3	AGE AT DIAGNOSIS: 31934

At the bottom, there is a green button labeled 'Predict Cancer Stage' and a link to 'Clear form'.

Figure 5.4: DriftNet input interface after data entry.

The left part of the page contains the methylation upload area. The system accepts a tab-separated text file containing CpG probe IDs and beta values. The interface also shows an example format so that users know how the file should be structured.

The right side contains the clinical feature form. This form includes the required fields used in the trained model, such as gender, vital status, race, ethnicity, sample type, tissue type, disease type, primary diagnosis, tissue or organ of origin, morphology code, age at diagnosis, prior treatment, and prior malignancy. The form

also includes a completion indicator so the user can confirm whether all required inputs are filled.

This input design reduces user error by keeping the format structured and by separating methylation input from clinical data input.

5.5 Output Design

After the user submits the required inputs, the backend runs the complete inference pipeline and returns the result. Figure 5.5 shows the output view of the system.

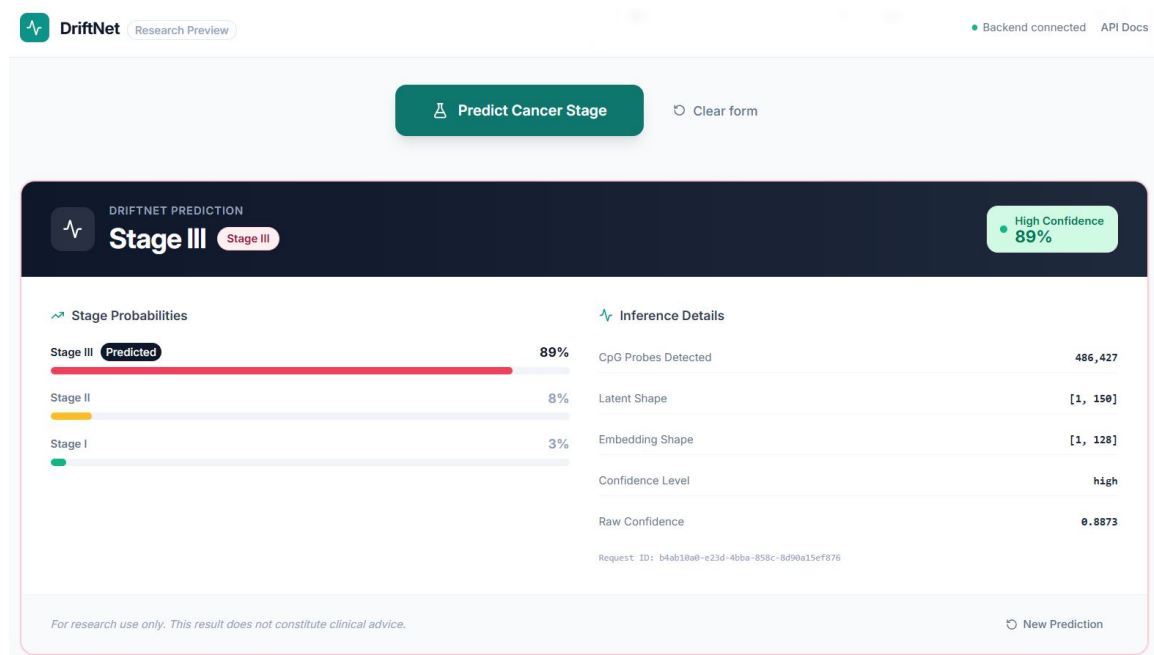


Figure 5.5: DriftNet prediction output interface.

The output panel is designed to present the result in a simple but informative way. The predicted stage is displayed clearly at the top. A confidence indicator is shown beside it so that the user can easily understand whether the prediction is strong or uncertain. Below that, the probability distribution across Stage I, Stage II, and Stage III is visualized using horizontal bars. This is useful because it shows not only the final class, but also how strongly the model favors each class.

The interface also presents several inference details, including the number of detected CpG probes, the latent feature shape, the embedding shape, and the raw

confidence value. These details improve transparency and help confirm that the file passed through the expected inference process.

5.6 Engineering Analysis and Design Decisions

The design of DriftNet was guided by several engineering considerations to ensure that the system works reliably as a research prototype.

5.6.1 Input Validation

The system requires one methylation file and all 13 clinical features. Therefore, validation is necessary before prediction begins. If the input is incomplete, the system should not continue. This reduces invalid inference requests and helps maintain reliable operation.

5.6.2 Feature Dimension Control

The trained pipeline depends on fixed input and output dimensions. The autoencoder must generate a 150-dimensional latent vector, and the contrastive model must generate a 128-dimensional embedding. If these shapes are not maintained, the final classifier will not work correctly. For this reason, feature dimension consistency is one of the core engineering requirements of the design.

5.6.3 Pipeline Consistency

A major design requirement is that the deployment pipeline must follow the same logic as the training pipeline. The same trained models, encoders, label mappings, and preprocessing rules are reused during inference. This ensures that the system behaves consistently with the way the model was originally trained and evaluated.

5.6.4 Usability and Error Reduction

Several interface design choices were made to reduce user mistakes:

- example file format is shown for the methylation input,
- uploaded file status is clearly displayed,

- many clinical fields are dropdown-based,
- related clinical variables are grouped into logical sections,
- a completion counter is displayed for the clinical form,
- the output is presented using both text and visual probability bars.

These design choices improve the usability of the system and make the process easier to follow.

5.7 Discussion of the Developed Design

The final design successfully solves the main system problem by connecting a complex machine-learning inference pipeline with a practical web-based user interface. The five figures presented in this chapter show the complete design flow of the platform: introduction, workflow explanation, main working page, input collection, and result presentation.

The homepage introduces the platform clearly. The workflow section explains the prediction logic. The main page provides the working interface. The input view shows how the system accepts the necessary user data, and the output view shows how the prediction result is reported back in an understandable form.

Together, these parts create a complete and organized research system that is understandable to users and faithful to the trained DriftNet pipeline.

5.8 Development Outcome

The final developed product is a functional web-based research prototype for pan-cancer stage prediction. It accepts a methylation file and clinical features, processes them through the trained DriftNet inference pipeline, and returns a predicted cancer stage together with supporting probability and confidence information.

Overall, the design demonstrates how a trained deep-learning model can be converted into a usable system. The platform is modular, readable, extendable, and suitable for research demonstration. However, it should still be treated as a research and decision-support prototype, not as a replacement for expert clinical judgement or a fully validated clinical tool.

Chapter 6

Plans and Deliverables

6.1 Timeline

This project is planned over a 12-month period, covering data preparation, model development, experimentation, and evaluation. The final months are allocated for result analysis, report writing, and thesis submission.

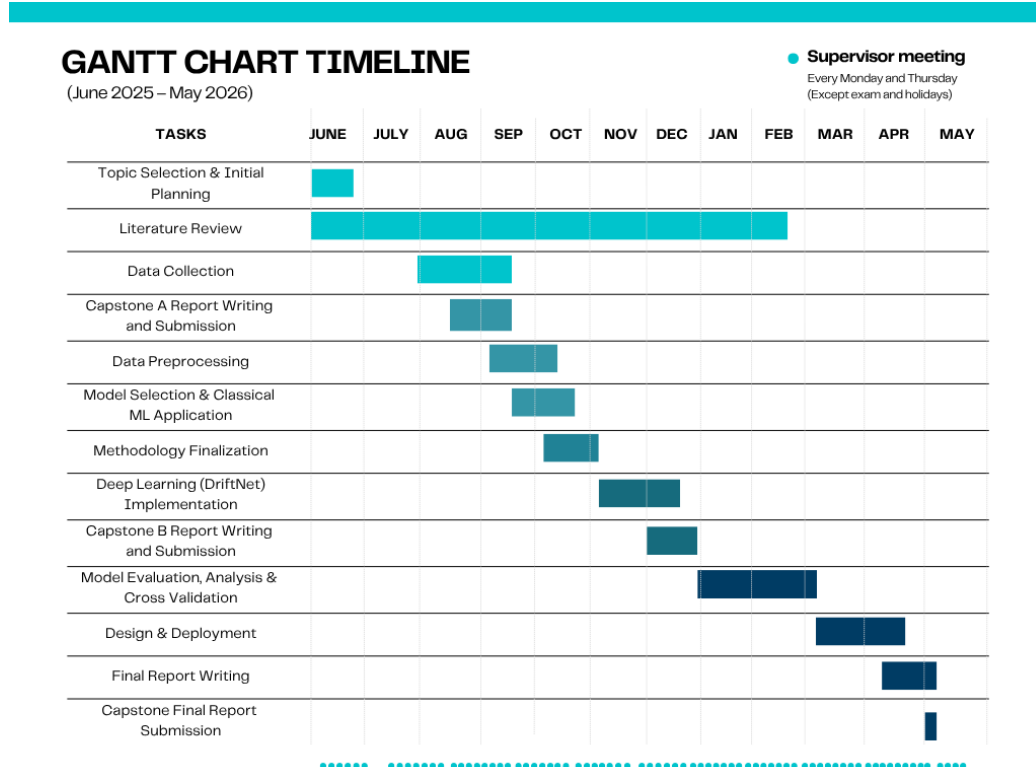


Figure 6.1: Project timeline for the 12-month DriftNet research study

6.2 Budget Planning

This project will run for 12 months and mainly uses computer-based work. The data used in this research are already available online for free, and all software tools are open source. However, some budget is still needed to support GPU computing, internet usage, cloud storage, and printing the final report.

Cloud GPU services are needed to train deep-learning models and to run experiments many times. Internet access is required to download large datasets and use online platforms. Cloud storage helps keep data, trained models, and results safely backed up. A small amount of money is also needed for printing and binding the project report. Overall, the budget is kept low and realistic, suitable for a student research project.

Table 6.1: Estimated budget planning for the 12-month DriftNet research project

Budget Item	Description	Estimated Cost (BDT)
Computing & GPU Resources	Cloud-based GPU usage for deep-learning model training and repeated experiments	24,000
Software Tools	Open-source tools (Python, TensorFlow, PyTorch, Scikit-learn)	0
Dataset Acquisition	TCGA methylation and clinical data (publicly available)	0
Internet Usage	Research-related internet usage for data download and cloud access	12,000
Cloud Storage & Backup	Online storage for datasets, trained models, and experiment results	6,000
Documentation & Printing	Printing, binding, and submission of reports	3,000
Total Estimated Budget		45,000

Chapter 7

Conclusion

Accurate cancer staging is important for treatment planning, but it remains challenging because cancer progression is complex and can vary across tumour types. This study investigated whether artificial intelligence can support pan-cancer Stage I/II/III prediction using DNA methylation-derived features and basic clinical information. Using 1,343 TCGA patients across four retained primary diagnosis groups, DriftNet achieved the best overall performance among all evaluated models. The final held-out test accuracy was 67.82%, with Cohen’s kappa of 0.5049, macro F1-score of 0.6442, and weighted F1-score of 0.6829.

DriftNet combines three main components: autoencoder-based methylation compression, supervised triplet-loss contrastive learning, and a dual-branch neural network for final stage classification. The autoencoder reduced the high-dimensional methylation profile into 150 latent features. The contrastive learning stage then transformed the pre-contrastive feature representation into a 128-dimensional stage-aware embedding. This embedding was combined with 13 clinical features to predict Stage I, Stage II, or Stage III.

A key finding of this study is that representation learning is valuable for high-dimensional biological data. Using the original full-dimensional methylation features achieved 56.13% accuracy, while the autoencoder-based representation with the DriftNet pipeline achieved 67.82% accuracy. This improvement shows that compact feature learning can reduce noise and support better generalisation when the number of patient samples is much smaller than the number of CpG features.

The final evaluation also showed that DriftNet performed better than both classical machine-learning baselines and paper-inspired deep-learning baselines. DriftNet ranked first overall, achieving the highest accuracy, Cohen’s kappa, macro

F1-score, weighted F1-score, and Stage III F1-score. The strongest classical baseline, Gradient Boosting, achieved 65.84% accuracy, while the strongest paper-inspired deep baseline, CMLV-AE + ANN, achieved 64.85% accuracy. DriftNet also achieved the highest Stage III F1-score of 0.5138, showing better recognition of the difficult advanced-stage class compared with the baseline models.

The model's stability was evaluated through repeated random-seed experiments and strict cross-validation. Across 20 random-seed runs, DriftNet achieved a mean test accuracy of 66.46% with a standard deviation of 0.87%. In strict 5-fold cross-validation, it achieved $65.67\% \pm 2.62\%$ accuracy. These results suggest that the model did not depend only on one favourable split and maintained reasonably stable performance across different evaluation settings.

However, this study also has limitations. Stage II remained the most difficult class, and many errors occurred between Stage II and Stage III. This is expected because adjacent cancer stages may share similar methylation patterns and may also require additional clinical information, such as tumour size, lymph-node involvement, and detailed pathological measurements. The model was also evaluated only on TCGA data, so external validation on independent cohorts is required before any clinical application.

Future work should include additional clinical and molecular data types, such as gene expression, copy-number variation, tumour size, lymph-node status, and survival-related variables. Future versions of DriftNet may also benefit from ordinal loss functions because cancer stages follow a natural order. Another useful direction is uncertainty estimation, which would allow the model to flag low-confidence cases for expert review. Although this study used tissue-based methylation data, the same general idea may later be adapted to liquid biopsy data as the technology improves.

Overall, DriftNet provides a reproducible research framework for pan-cancer Stage I/II/III prediction using autoencoder-derived methylation features, contrastive learning, and clinical variables. While it is not ready to replace clinical staging, it shows that stage-aware representation learning can improve cancer stage prediction and can serve as a useful foundation for future decision-support research.

7.1 Social, Legal, Ethical, and Environmental Issues

This research is based entirely on retrospective, de-identified data from The Cancer Genome Atlas (TCGA), which follows established ethical guidelines for patient consent and data usage. Since there was no direct patient interaction or intervention, the study carries lower direct clinical risk. However, responsible use is still important because AI-based staging predictions could influence medical interpretation if used without proper validation.

From an ethical perspective, the use of artificial intelligence in cancer staging raises important considerations related to transparency, reliability, and clinical accountability. DriftNet is designed as a research decision-support prototype rather than a replacement for clinicians. Its predictions should be interpreted together with expert medical judgement, not used as independent clinical decisions. The observed ambiguity between Stage II and Stage III also highlights the need for uncertainty-aware deployment, where low-confidence or borderline predictions can be reviewed more carefully by experts.

From a legal perspective, the system should not be used in real clinical workflows without external validation, institutional approval, and appropriate regulatory review. Any future deployment would need clear responsibility for how predictions are interpreted and used. Data privacy is also important, especially if future versions use hospital or patient-specific data. Secure data handling, access control, and proper anonymisation would be required.

No major environmental concern arises directly from this study because it relies on computational analysis of existing digital datasets and does not involve laboratory experiments, chemical usage, or physical biological samples. However, deep-learning experiments still require computing resources, so future work should consider efficient training, model reuse, and responsible use of computational infrastructure.

7.2 Expected Final Product

The expected final product of this project is a complete deep-learning research system named DriftNet. The system is designed to predict cancer stage using

autoencoder-derived DNA methylation features together with basic clinical information. It works as a pan-cancer model, meaning one unified model is used across the retained cancer diagnosis groups instead of building a separate model for each cancer type.

In addition to the trained model, a web-based application will be developed to demonstrate the practical use of the system. Through this interface, users will be able to provide the required patient features and receive a predicted cancer stage in a simple and understandable format. The web application is intended mainly as a prototype to show how the trained model can be accessed, tested, and interpreted in a user-friendly way.

The final product will include the trained DriftNet model, the web application, source code, and supporting documentation. These outputs will make the system reproducible and extendable for future research. Overall, DriftNet is intended to serve as a research tool and prototype decision-support system. It may support future cancer-staging research, but it should not be used for real clinical decision-making without external validation, expert review, and proper regulatory approval.

References

- Amanzholova, A., & Coşkun, A. (2024). Enhancing cancer stage prediction through hybrid deep neural networks: a comparative study. *Frontiers in big Data*, 7, 1359703.
- Baylin, S. B., & Jones, P. A. (2011). A decade of exploring the cancer epigenome—biological and translational implications. *Nature Reviews Cancer*, 11(10), 726–734.
- Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8), 1798–1828.
- Brierley, J. D., Gospodarowicz, M. K., & Wittekind, C. (Eds.). (2017). *Tnm classification of malignant tumours* (8th ed.). Oxford, UK; Hoboken, NJ: Wiley-Blackwell.
- Cheerla, A., & Gevaert, O. (2019). Deep learning with multimodal representation for pancancer prognosis prediction. *Bioinformatics*, 35(14), i446–i454.
- Cruz, J. A., & Wishart, D. S. (2006). Applications of machine learning in cancer prediction and prognosis. *Cancer informatics*, 2, 117693510600200030.
- Edge, S. B., American Joint Committee on Cancer, A. C. S., et al. (2010). *Ajcc cancer staging handbook: from the ajcc cancer staging manual* (Vol. 19). Springer.
- Esteller, M. (2008). Epigenetics in cancer. *New England Journal of Medicine*, 358(11), 1148–1159.
- Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning* (pp. 1126–1135).
- Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J., & Greenspan, H. (2018). Synthetic data augmentation using gan for improved liver lesion classification. In *2018 ieee 15th international symposium on biomedical imaging (isbi 2018)* (pp. 289–293).

- Gomari, D. P., Schweickart, A., Cerchietti, L., Paietta, E., Fernandez, H., Al-Amin, H., ... Krumsiek, J. (2022). Variational autoencoders learn transferrable representations of metabolomics data. *Communications Biology*, 5(1), 645.
- Gurcan, F., & Soylu, A. (2024). Learning from imbalanced data: integration of advanced resampling techniques and machine learning models for enhanced cancer diagnosis and prognosis. *Cancers*, 16(19), 3417.
- Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar), 1157–1182.
- Hao, X., Luo, H., Krawczyk, M., Wei, W., Wang, W., Wang, J., ... others (2017). Dna methylation markers for diagnosis and prognosis of common cancers. *Proceedings of the National Academy of Sciences*, 114(28), 7414–7419.
- Hoadley, K. A., Yau, C., Hinoue, T., Wolf, D. M., Lazar, A. J., Drill, E., ... others (2018). Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell*, 173(2), 291–304.
- Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., & Xu, W. (2018). Applications of support vector machine (svm) learning in cancer genomics. *Cancer genomics & proteomics*, 15(1), 41–51.
- Huang, Z., Zhan, X., Xiang, S., Johnson, T. S., Helm, B., Yu, C. Y., ... others (2019). Salmon: survival analysis learning with multi-omics neural networks on breast cancer. *Frontiers in genetics*, 10, 166.
- Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning* (pp. 448–456).
- Irizarry, R. A., Ladd-Acosta, C., Wen, B., Wu, Z., Montano, C., Onyango, P., ... Feinberg, A. P. (2009, February). The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nature Genetics*, 41(2), 178–186.
- Ivan, J., Patricia, G., & Agustriawan, D. (2021). In silico study of cancer stage-specific dna methylation pattern in white breast cancer patients based on tcga dataset. *Computational Biology and Chemistry*, 92, 107498.
- Kohavi, R., et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai* (Vol. 14, pp. 1137–1145).
- Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, 13, 8–17.

- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, 521(7553), 436–444.
- Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision (iccv)* (pp. 2980–2988). Venice, Italy: IEEE.
- Liu, T. Y., & Feng, J. (2021). Triplet contrastive learning for brain tumor classification. *arXiv preprint arXiv:2108.03611*.
- Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., & Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12), 1053–1058.
- McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3), 276–282.
- Munquad, S., Dash, B. K., Sengupta, S., Singh, V., Ushakiran, Y., & Das, A. B. (2025). Biologically explainable multi-omics feature demonstrates greater learning potential by identifying tissue of origin, stages, and subtypes for pan-cancer classification. *Scientific Reports*, 15(1), 40627.
- Nair, A. (2024). *Multimodal deep learning models for predicting cancer stage: A pan-cancer analysis* (Honors Thesis). Brown University, Providence, RI.
- Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10), 1345–1359.
- Powers, D. M. (2020). Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*.
- Prechelt, L. (2002). Early stopping-but when? In *Neural networks: Tricks of the trade* (pp. 55–69). Springer.
- Rappoport, N., & Shamir, R. (2018). Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic acids research*, 46(20), 10546–10562.
- Ribeiro, M. T., Singh, S., & 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* (pp. 1135–1144). San Francisco, CA: ACM.
- Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition (cvpr)* (pp. 815–823). Boston, MA: IEEE.
- Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4), 427–437.

- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1), 1929–1958.
- Wei, R., & Mahmood, A. (2020). Recent advances in variational autoencoders with representation learning for biomedical informatics: A survey. *Ieee Access*, 9, 4939–4956.
- Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., ... Stuart, J. M. (2013). The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10), 1113–1120.
- Weisenberger, D. J., et al. (2014). Characterizing dna methylation alterations from the cancer genome atlas. *The Journal of clinical investigation*, 124(1), 17–23.
- Widschwendter, M., & Jones, P. A. (2002). Dna methylation and breast carcinogenesis. *Oncogene*, 21(35), 5462–5482.
- Wong, T.-T., & Yeh, P.-Y. (2019). Reliable accuracy estimates from k-fold cross validation. *IEEE Transactions on Knowledge and Data Engineering*, 32(8), 1586–1594.
- Wu, Y., Lin, K., Huang, H., Li, F., & Xu, B. (2025). Breast ultrasound image classification based on dynamic adversarial adaptive network with triplet loss. In *Proceedings of the 2025 5th international conference on neural networks, information and communication engineering (nnice)* (pp. 431–435). Beijing, China: IEEE.
- Yeung, M., Sala, E., Schönlieb, C.-B., & Rundo, L. (2022). Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation. *Computerized Medical Imaging and Graphics*, 95, 102026.
- Zhang, X., Zhang, J., Sun, K., Yang, X., Dai, C., & Guo, Y. (2019). Integrated multi-omics analysis using variational autoencoders: application to pan-cancer classification. In *2019 ieee international conference on bioinformatics and biomedicine (bibt)* (pp. 765–769).
- Zhao, C., Liu, A., Zhang, X., Cao, X., Ding, Z., Sha, Q., ... Zhou, W. (2023). Clcls: Cross-omics linked embedding with contrastive learning and self attention for multi-omics integration with incomplete multi-omics data. *Research Square*, rs-3.