Appendix A

Appendix A describes the preprocessing steps we performed for the dataset used in our study.

breast_cancer (UCTH)

The target class (*Diagnosis Result*) was mapped to 0 (Benign) and 1 (Malignant). Initially, the unknown values were marked as #, which were changed to missing (NaN) values. Rows containing missing values were then dropped from the dataset, as few of them were there. The column S/N was also dropped as it does not provide any value to the model's training since it is a unique identifier.

cvd_uae

The *StudyID* column was dropped as it is a unique identifier, which does not provide any meaningful information to the training of the models.

haberman

The *Survival* column was stripped of leading/trailing whitespace characters, and its string values were mapped to numerical representations, with 'negative' encoded as 0 and 'positive' as 1.

hcc_data

Numerical columns with missing values were imputed using median values, which were calculated separately for each class (positive and negative).

mammographic

Initially, the unknown values were marked as ?, which were changed to missing (NaN) values. Rows containing missing values were then dropped from the dataset, as few of them were there.

wisconsin

The target class (*Class*) was mapped to 0 (for benign (2)) and to 1 (for malignant (4)). Rows containing missing values were then dropped from the dataset, as few of them were there.

obesity

The original dataset had multiple classes for categorizing patients based on weight status: insufficient, normal, overweight, and obese. This was transformed into a binary classification task, with the normal weight class labeled as 0 (healthy) and the remaining classes (insufficient, overweight, and obese) consolidated and labeled as 1 (unhealthy).

hepatitis

The target variable (*Class*) was mapped to binary classes, with two mapped to 0 and 1 mapped to 1. For columns with fewer missing values (*STEROID*, *FATIGUE*, *MALAISE*, *ANOREXIA*, *LIVER_BIG*, *LIVER_FIRM*, *SPLEEN_PALPABLE*, *SPIDERS*, *ASCITES*, *VARICES*, *BILIRUBIN*, *SGOT*, *ALBUMIN*), the missing values were imputed with the median value of the corresponding class (0 or 1) for that column. Columns with a high proportion of missing values (*ALK_PHOSPHATE*, *PROTIME*) were dropped from the dataset.

parkinsons

The *name* column was dropped as it is a unique identifier, which does not provide any meaningful information to the training of the models.

arrhythmia

The dataset's unknown values represented by '?' were replaced with missing values (NaN). The column names were then updated to more meaningful names, such as *Age*, *Sex*, *Height*, *Weight*, *QRSduration*, and so on, including various measurements related to different ECG channels (DI, DII, DIII, AVR, AVL, AVF, V1, V2, V3, V4, V5, V6). The *J* column was dropped from the dataset, as it contained zeros only for both classes, which does not provide any value but may add to the computational cost during training. Rows containing missing values were then dropped from the dataset, as few of them were there.

ilpd

The column names were retrieved from the repository and assigned programmatically, as the dataset did not include column names by default. The *Selector* column containing the target variable was mapped to binary classes, with two mapped to 0 and 1 mapped to 1. Rows containing missing values were then dropped from the dataset, as few of them were there.

darwin

The *ID* column, which contained unique identifiers, was dropped from the dataset. The target variable *Class* was mapped to binary classes, with 'P' mapped to 1 and 'H' mapped to 0.

lymphography

In this dataset, the positive cases (classes 2, 3, and 4) significantly outnumbered the healthy cases (class 1). Therefore, the preprocessing treated the positive cases as the negative class and vice versa. The target variable *Class* was mapped such that 1 was assigned to 0 (healthy), and 2, 3, and 4 were assigned to 1 (positive cases treated as negative). Subsequently, the classes were flipped, with 0 mapped to 1 and 1 mapped to 0, effectively swapping the positive and negative classes.

breast_cancer_coimbra

The target variable *Class* was mapped to binary classes, with ’1’ mapped to 0 and ’2’ mapped to 1.

cirrhosis

In this dataset, the preprocessing treated the ’D’ (death) event as the positive case, while ’C’ (censored) and ’CL’ (censored after live transplant) were considered negative cases. The objective was to predict whether the patient survived after the trial or not. The target variable *Class* was mapped accordingly, with ’D’ assigned to 1 (positive) and ’C’ and ’CL’ assigned to 0 (negative). The *ID* column was dropped as it contained unique identifiers, which does not provide any meaningful information to the training of the models. The *Age* column was converted from days to years by dividing by 365 and rounding to the nearest integer. Missing values were handled separately for positive and negative cases. For categorical columns (*Drug*, *Ascites*, *Hepatomegaly*, *Spiders*), missing values were imputed with the mode of the corresponding class. For numerical columns with fewer missing entries (*Platelets*, *Prothrombin*, *Stage*), missing values were imputed with the median of the corresponding class. Columns with substantial missing data (*Cholesterol*, *Copper*, *Tryglicerides*) had their missing values imputed with the median of the entire column. Finally, the columns *Alk_Phos* and *SGOT* were dropped from the dataset as they contained missing values more than 50%.

bone_marrow

Rows containing missing values were then dropped from the dataset, as there were not many of them.

thyroid

The target class (*Class*) was mapped to 0 (for (healthy(1)) and to 1 (suffers from hyperthyroidism (2) or hypothyroidism (3)).

Appendix B

Appendix B outlines the errors we faced during the training of our experiment, where we show the dataset, model, and the error.

Table 1. Errors during training

Dataset	Model	Error
hcc_data	CD	Input contains NaN.
hcc_data	AutoEncoder	Expected state_dict to be dict-like, got <class ’NoneType’>.
hcc_data	MO_GAAL	Unexpected result of ‘predict_function‘ (Empty batch_outputs). Please use ‘Model.compile(..., run_eagerly=True)‘, or ‘tf.config.run_functions_eagerly(True)‘ for more information of where went wrong, or file a issue/bug to ‘tf.keras‘.
spectfheart	CD	Input contains NaN.
spectfheart	MO_GAAL	Unexpected result of ‘predict_function‘ (Empty batch_outputs). Please use ‘Model.compile(..., run_eagerly=True)‘, or ‘tf.config.run_functions_eagerly(True)‘ for more information of where went wrong, or file a issue/bug to ‘tf.keras‘.
parkinsons	CD	Input contains NaN.
parkinsons	MO_GAAL	Unexpected result of ‘predict_function‘ (Empty batch_outputs). Please use ‘Model.compile(..., run_eagerly=True)‘, or ‘tf.config.run_functions_eagerly(True)‘ for more information of where went wrong, or file a issue/bug to ‘tf.keras‘.
arrhythmia	AutoEncoder	Expected state_dict to be dict-like, got <class ’NoneType’>.
darwin	GMM	Fitting the mixture model failed because some components have ill-defined empirical covariance (for instance caused by singleton or collapsed samples). Try to decrease the number of components, or increase reg_covar.
darwin	CD	Input contains NaN.
darwin	CD	Input contains NaN.
breast_cancer _coimbra	MO_GAAL	Unexpected result of ‘predict_function‘ (Empty batch_outputs). Please use ‘Model.compile(..., run_eagerly=True)‘, or ‘tf.config.run_functions_eagerly(True)‘ for more information of where went wrong, or file a issue/bug to ‘tf.keras‘.
cirrhosis	AutoEncoder	Expected state_dict to be dict-like, got <class ’NoneType’>.
cervical_cancer	CD	Input contains NaN.
cervical_cancer	MO_GAAL	Unexpected result of ‘predict_function‘ (Empty batch_outputs). Please use ‘Model.compile(..., run_eagerly=True)‘, or ‘tf.config.run_functions_eagerly(True)‘ for more information of where went wrong, or file a issue/bug to ‘tf.keras‘.
bone_marrow	CD	Input contains NaN.