Attention-Enhanced Deep Learning for Cardiovascular Disease Prediction

Omkar Badadale

CIS 602: Fundamentals of Deep Learning

University of Massachusetts Dartmouth

10/19/2025

Project Repository:

Contact:

**ABSTRACT**

Cardiovascular disease (CVD) remains the leading cause of mortality worldwide, accounting for 17.9 million deaths annually. Early risk prediction is crucial for preventive intervention, yet traditional machine learning models often struggle with severe class imbalance and lack clinical interpretability. This study presents a novel deep learning approach for cardiovascular disease prediction using the CDC Behavioral Risk Factor Surveillance System (BRFSS) dataset comprising 253,680 patient records. We implemented three novel components: (1) a TabNet-inspired attention mechanism for interpretable feature selection, (2) Focal Loss optimization to address severe class imbalance (9.62:1 ratio), and (3) Cosine Annealing Warm Restarts for stable training dynamics. Our attention-enhanced model achieved 81.3% recall (sensitivity) compared to 78.1% for the baseline Multi-Layer Perceptron, representing a 3.2 percentage point improvement in detecting at-risk patients. ROC-AUC performance was maintained at 84.4%, demonstrating strong discriminative ability. Feature importance analysis revealed that the model prioritized clinically validated risk factors including prior stroke (23% attention weight), high blood pressure (11%), and high cholesterol (8%), validating the model's learned representations. The results demonstrate that attention mechanisms combined with specialized loss functions can significantly improve recall in imbalanced healthcare datasets while maintaining interpretability—a critical requirement for clinical deployment.

**Keywords:** Deep Learning, Cardiovascular Disease, Attention Mechanism, Focal Loss, Class Imbalance, Healthcare AI, TabNet, PyTorch

**1. INTRODUCTION**

**1.1 Background and Motivation**

Cardiovascular disease (CVD) encompasses a range of heart and blood vessel disorders, including coronary artery disease, heart attacks, and strokes. According to the World Health Organization, CVD is responsible for approximately 32% of all global deaths, making it the leading cause of mortality worldwide. In the United States alone, one person dies every 36 seconds from cardiovascular disease, costing the healthcare system approximately $219 billion annually in direct medical costs and lost productivity.

Early identification of at-risk individuals is paramount for implementing preventive interventions such as lifestyle modifications, medication, and increased monitoring. Traditional risk assessment tools like the Framingham Risk Score and ASCVD (Atherosclerotic Cardiovascular Disease) calculator rely on linear models with predetermined weights for clinical variables. While these tools have proven valuable, they have limitations:

1. Fixed coefficients that may not capture complex non-linear interactions between risk factors
2. Population-specific calibration that may not generalize across diverse demographics
3. Limited feature sets that exclude potentially informative behavioral and socioeconomic factors
4. Lack of adaptability to incorporate new biomarkers or risk factors as they emerge

Machine learning and deep learning approaches offer the potential to overcome these limitations by learning complex patterns directly from data. However, healthcare datasets present unique challenges:

* Severe class imbalance: Disease prevalence is typically low (5-15%), leading to imbalanced datasets
* Interpretability requirements: Clinical adoption requires understanding why a model makes predictions
* Cost asymmetry: False negatives (missing disease) are far more costly than false positives
* Regulatory constraints: Medical AI systems require rigorous validation and transparency

**1.2 Problem Statement**

This project addresses the challenge of predicting cardiovascular disease from tabular health indicators while handling severe class imbalance and maintaining clinical interpretability. Specifically, we aim to:

1. Develop a deep learning model that achieves high recall (sensitivity) to minimize false negatives
2. Implement attention mechanisms to provide interpretable feature importance scores
3. Address class imbalance through specialized loss functions
4. Compare performance against a baseline model to quantify improvements

**1.3 Related Work**

**Deep Learning for Tabular Data:**

While deep learning has achieved remarkable success in computer vision and natural language processing, its application to tabular data has been more challenging. Traditional gradient boosting methods (XGBoost, LightGBM) have often outperformed neural networks on structured data. However, recent architectures like TabNet (Arik & Pfister, 2021) have demonstrated that attention mechanisms can make deep learning competitive for tabular tasks by learning which features to attend to at each decision step.

**Class Imbalance in Healthcare:**

Medical datasets frequently exhibit severe class imbalance, with disease prevalence ranging from 1-15%. Standard approaches include resampling techniques (SMOTE, ADASYN), cost-sensitive learning (class weighting), and specialized loss functions. Focal Loss, originally developed for object detection (Lin et al., 2017), has shown promise in medical applications by down-weighting easy examples and focusing learning on hard cases.

**Cardiovascular Disease Prediction:**

Previous machine learning studies on CVD prediction have employed various techniques including logistic regression, random forests, support vector machines, and neural networks. Recent deep learning approaches have explored recurrent neural networks for temporal risk modeling and convolutional networks for medical imaging. However, most tabular CVD prediction studies have not incorporated attention mechanisms for interpretability or systematically addressed class imbalance with Focal Loss.

**1.4 Contributions**

This work makes the following contributions:

1. Novel architecture: Implementation of TabNet-inspired attention mechanisms for cardiovascular disease prediction from tabular data
2. Class imbalance handling: Systematic application and tuning of Focal Loss (α=0.25, γ=2.0) for healthcare screening
3. Training optimization: Integration of Cosine Annealing Warm Restarts for stable convergence
4. Interpretability analysis: Extraction and validation of learned feature importance against clinical knowledge
5. Comprehensive evaluation: Comparison using healthcare-relevant metrics (recall, ROC-AUC, clinical decision curves)

**1.5 Report Organization**

The remainder of this report is organized as follows: Section 2 describes the dataset, preprocessing pipeline, and experimental methodology. Section 3 presents the model architectures, novel components, and training procedures. Section 4 reports experimental results with comprehensive visualizations. Section 5 discusses findings, clinical implications, and limitations. Section 6 concludes with key takeaways and future work directions.

**2. METHODOLOGY**

**2.1 Dataset Description**

**Dataset Source:**

This study utilizes the CDC Behavioral Risk Factor Surveillance System (BRFSS) 2015 dataset, publicly available through Kaggle. The BRFSS is an annual telephone survey conducted by the Centers for Disease Control and Prevention (CDC) to collect data on health-related risk behaviors, chronic health conditions, and use of preventive services among U.S. residents.

**Dataset Characteristics:**

* Total samples: 253,680 individual patient records
* Number of features: 21 clinical and behavioral indicators
* Target variable: Binary classification (HeartDiseaseorAttack: 0 = No disease, 1 = Disease or heart attack)
* Class distribution: 229,787 negative cases (90.6%), 23,893 positive cases (9.4%)
* Imbalance ratio: 9.62:1
* Missing values: None (preprocessed dataset)
* Data type: All features are numerical (binary or continuous)

**Features Overview:**

*Clinical Risk Factors:*

* HighBP: High blood pressure diagnosis (binary)
* HighChol: High cholesterol diagnosis (binary)
* CholCheck: Cholesterol check in past 5 years (binary)
* BMI: Body Mass Index (continuous, range: 12-98)
* Stroke: History of stroke (binary)
* Diabetes: Diabetes status (0=no, 1=prediabetes, 2=diabetes)

*Lifestyle and Behavioral Factors:*

* Smoker: Smoked at least 100 cigarettes lifetime (binary)
* PhysActivity: Physical activity in past 30 days (binary)
* Fruits: Consume fruit 1+ times per day (binary)
* Veggies: Consume vegetables 1+ times per day (binary)
* HvyAlcoholConsump: Heavy alcohol consumption (binary)

*Healthcare Access:*

* AnyHealthcare: Has any healthcare coverage (binary)
* NoDocbcCost: Could not see doctor due to cost (binary)

*Health Status:*

* GenHlth: General health (1=excellent to 5=poor)
* MentHlth: Days of poor mental health in past 30 days (0-30)
* PhysHlth: Days of poor physical health in past 30 days (0-30)
* DiffWalk: Difficulty walking or climbing stairs (binary)

*Demographics:*

* Sex: Biological sex (0=female, 1=male)
* Age: Age category (1=18-24 to 13=80+)
* Education: Education level (1=never attended to 6=college graduate)
* Income: Annual household income (1=<$10k to 8=$75k+)

**[INSERT VISUALIZATION HERE: Figure 1 - Class Distribution]**  
*Caption: Distribution of cardiovascular disease cases in the BRFSS 2015 dataset. The severe imbalance (9.62:1 ratio) motivated the use of Focal Loss and SMOTE oversampling.*

**2.2 Exploratory Data Analysis**

**Class Imbalance Analysis:**

The dataset exhibits severe class imbalance with only 9.4% of samples representing positive cases (heart disease or heart attack). This imbalance is characteristic of screening datasets where disease prevalence is naturally low. Standard machine learning approaches optimizing for accuracy would achieve 90.6% accuracy by simply predicting "no disease" for all cases—a clinically useless model that misses every at-risk patient. This finding underscores the necessity of specialized techniques for imbalanced classification.

**[INSERT VISUALIZATION HERE: Figure 2 - Feature Distributions by Class]**  
*Caption: Distribution of key features stratified by disease status. Notable separation is observed for Age, BMI, GenHlth, HighBP, and HighChol, indicating their discriminative power.*

**Feature Distribution Analysis:**

Visual inspection of feature distributions revealed that several variables show clear separation between positive and negative classes:

* Age: Disease cases skew toward older age groups (categories 8-13)
* BMI: Higher body mass index associated with increased disease prevalence
* GenHlth: Poor general health (categories 4-5) strongly associated with disease
* HighBP and HighChol: Prevalence of hypertension and hypercholesterolemia significantly elevated in disease group
* PhysActivity: Lower physical activity levels observed in disease cases

**Correlation Analysis:**

Pearson correlation analysis was conducted to identify potential multicollinearity and understand feature relationships with the target variable.

**[INSERT VISUALIZATION HERE: Figure 3 - Correlation Matrix Heatmap]**  
*Caption: Correlation matrix of all features. No feature pairs exceeded |r| = 0.7, indicating absence of severe multicollinearity.*

**Key findings:**

* No severe multicollinearity: No feature pairs exhibited correlation > 0.7, allowing retention of all 21 features
* Top correlates with target:
  + GenHlth (General Health): r = +0.258
  + Age: r = +0.222
  + DiffWalk: r = +0.213
  + HighBP: r = +0.209
  + Stroke: r = +0.203
* Protective factors: Physical activity (r = -0.11), Education (r = -0.08), and Income (r = -0.10) showed negative correlations, suggesting protective effects

These moderate correlations (0.2-0.3 range) are typical in healthcare datasets and justify the use of ensemble learning approaches like deep neural networks that can capture complex non-linear interactions.

**2.3 Data Preprocessing**

A systematic preprocessing pipeline was implemented to prepare the data for deep learning:

**2.3.1 Feature Categorization**

Features were categorized based on their value ranges:

* Binary features (13): HighBP, HighChol, CholCheck, Smoker, Stroke, PhysActivity, Fruits, Veggies, HvyAlcoholConsump, AnyHealthcare, NoDocbcCost, DiffWalk, Sex
* Continuous features (8): BMI, Diabetes, GenHlth, MentHlth, PhysHlth, Age, Education, Income

**2.3.2 Train/Validation/Test Split**

Data was split using stratified sampling to preserve class distribution:

* Training set: 177,677 samples (70%)
* Validation set: 37,951 samples (15%)
* Test set: 38,052 samples (15%)

Stratification ensured that each split maintained the 90.6%/9.4% class ratio, preventing biased evaluation. The validation set was used for hyperparameter tuning and early stopping, while the test set was held out for final model evaluation.

**2.3.3 Feature Scaling**

To prevent features with larger magnitude ranges from dominating the learning process, StandardScaler normalization was applied:

* Continuous features: Transformed to zero mean and unit variance: x' = (x - μ) / σ
* Binary features: Left unchanged (already on 0/1 scale)

**Critical consideration:** The scaler was fitted exclusively on training data to prevent data leakage. The same transformation was then applied to validation and test sets.

Verification:

BMI before scaling: mean=28.39, std=6.62

BMI after scaling: mean=0.00, std=1.00 ✓

```

\*\*2.3.4 SMOTE Oversampling\*\*

Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training set to address class imbalance at the data level:

\*\*Before SMOTE:\*\*

- Class 0 (No disease): 160,942 samples (90.58%)

- Class 1 (Disease): 16,735 samples (9.42%)

\*\*After SMOTE:\*\*

- Class 0: 160,942 samples (50.00%)

- Class 1: 160,942 samples (50.00%)

- Total training samples: 321,884 (144,207 synthetic samples generated)

SMOTE generates synthetic samples by interpolating between existing minority class examples in feature space. Using k=5 nearest neighbors, the algorithm creates new samples along the line segments connecting minority class instances.

\*\*Important:\*\* SMOTE was applied only to the training set. Validation and test sets retained their natural imbalanced distributions to provide realistic performance estimates.

\*\*2.3.5 PyTorch Tensor Conversion\*\*

Preprocessed data was converted to PyTorch tensors:

- Data type: Float32 for features, Long (Int64) for labels

- Device: Moved to Apple MPS (Metal Performance Shaders) for GPU acceleration

- DataLoader configuration:

- Batch size: 512 (optimized for M4 Pro)

- Shuffle: True for training, False for validation/test

- num\_workers: 0 (MPS compatibility)

Final tensor shapes:

- Training: X=[321,884 × 21], y=[321,884]

- Validation: X=[37,951 × 21], y=[37,951]

- Test: X=[38,052 × 21], y=[38,052]

### 2.4 Evaluation Metrics

Given the class imbalance and healthcare context, we employed multiple evaluation metrics beyond accuracy:

\*\*Primary Metrics:\*\*

1. \*\*Recall (Sensitivity / True Positive Rate):\*\*

- Formula: TPR = TP / (TP + FN)

- Most critical metric for healthcare screening

- Measures the proportion of actual disease cases correctly identified

- High recall minimizes false negatives (missed diagnoses)

2. \*\*ROC-AUC (Area Under Receiver Operating Characteristic Curve):\*\*

- Measures discriminative ability across all classification thresholds

- Interpretation: Probability that model ranks a random positive case higher than random negative case

- Range: 0.5 (random) to 1.0 (perfect)

- Robust to class imbalance

\*\*Secondary Metrics:\*\*

3. \*\*Precision (Positive Predictive Value):\*\*

- Formula: PPV = TP / (TP + FP)

- Proportion of predicted positives that are actually positive

- Indicates false alarm rate

4. \*\*F1-Score:\*\*

- Harmonic mean of precision and recall: 2 × (Precision × Recall) / (Precision + Recall)

- Balanced metric for imbalanced datasets

5. \*\*Accuracy:\*\*

- Overall correctness: (TP + TN) / (TP + TN + FP + FN)

- Less informative for imbalanced datasets

6. \*\*PR-AUC (Precision-Recall Area Under Curve):\*\*

- More informative than ROC-AUC for severely imbalanced datasets

- Focuses on performance in the minority class region

\*\*Clinical Evaluation:\*\*

7. \*\*Confusion Matrix:\*\*

- Visual breakdown of TP, TN, FP, FN

- Enables calculation of clinical metrics (sensitivity, specificity, PPV, NPV)

8. \*\*Decision Curve Analysis:\*\*

- Evaluates net benefit of model-guided decisions across threshold range

- Compares model to "treat all" and "treat none" strategies

- Clinically actionable metric

\*\*Justification:\*\*

In cardiovascular disease screening, recall is paramount. Missing an at-risk patient (false negative) can lead to preventable heart attacks or deaths, whereas false positives result in additional testing—an acceptable trade-off. Therefore, our optimization strategy prioritized maximizing recall while maintaining reasonable ROC-AUC.

---

## 3. MODEL ARCHITECTURE AND NOVEL COMPONENTS

### 3.1 Baseline Model: Multi-Layer Perceptron

To establish performance benchmarks, we first implemented a standard feedforward neural network.

\*\*Architecture:\*\*

```

Input Layer: 21 features

Hidden Layer 1: 128 neurons, ReLU activation, Dropout(0.3)

Hidden Layer 2: 64 neurons, ReLU activation, Dropout(0.2)

Hidden Layer 3: 32 neurons, ReLU activation

Output Layer: 2 neurons (binary classification logits)

```

\*\*Configuration:\*\*

- Total parameters: 13,218

- Loss function: CrossEntropyLoss (standard)

- Optimizer: Adam (lr=0.001, weight\_decay=0.01)

- Regularization: Dropout layers (0.3, 0.2)

- Training epochs: 20 (with early stopping, patience=10)

- Batch size: 512

\*\*Rationale:\*\*

This architecture serves as a strong baseline using established techniques. The three-hidden-layer structure provides sufficient capacity for learning complex patterns, while dropout prevents overfitting. CrossEntropyLoss is the standard choice for classification tasks.

### 3.2 Novel Component #1: Attention Mechanism

\*\*Motivation:\*\*

Traditional neural networks treat all input features equally, applying learned transformations without explicit feature selection. In healthcare, understanding which features drive predictions is critical for clinical trust and regulatory approval. Attention mechanisms address this by learning to dynamically weight feature importance.

\*\*TabNet-Inspired Architecture:\*\*

We implemented a simplified attention mechanism inspired by TabNet (Arik & Pfister, 2021), which introduced sequential attention for tabular data. Our AttentionBlock learns a soft feature selection mask:

\*\*Attention Block Architecture:\*\*

```

AttentionBlock:

Linear(21 → 64)

BatchNorm1d(64)

ReLU()

Dropout(0.2)

Linear(64 → 21)

Softmax(dim=1) # Outputs attention weights summing to 1.0

```

\*\*Mechanism:\*\*

1. Input features x ∈ R^21 are passed through a small neural network

2. Softmax normalization ensures attention weights α ∈ R^21 sum to 1.0

3. Element-wise multiplication creates attended features: x' = x ⊙ α

4. Attended features are processed by the main classification network

\*\*Mathematical Formulation:\*\*

```

α = Softmax(W₂ · ReLU(BN(W₁ · x)))

x' = x ⊙ α

```

Where:

- W₁ ∈ R^{64×21}, W₂ ∈ R^{21×64} are learned weight matrices

- BN is batch normalization

- ⊙ denotes element-wise multiplication

\*\*Full Model Architecture:\*\*

```

AttentionBlock (feature selection)

↓

Main Processing Network:

Linear(21 → 128), BatchNorm1d(128), ReLU, Dropout(0.3)

Linear(128 → 64), BatchNorm1d(64), ReLU, Dropout(0.2)

Linear(64 → 32), BatchNorm1d(32), ReLU

Linear(32 → 2) # Output logits

```

\*\*Total parameters:\*\* 18,690 (+41.4% vs baseline)

\*\*Interpretability Benefit:\*\*

The attention weights α can be extracted and averaged across the test set to reveal which features the model considers most important globally, or examined per-sample for instance-level explanations.

\*\*Reference:\*\* Arik, S. Ö., & Pfister, T. (2021). TabNet: Attentive interpretable tabular learning. \*AAAI Conference on Artificial Intelligence\*, 35(8), 6679-6687.

### 3.3 Novel Component #2: Focal Loss

\*\*Motivation:\*\*

Standard CrossEntropyLoss treats all samples equally, which is problematic for imbalanced datasets. While SMOTE addresses imbalance at the data level, algorithmic solutions can provide additional benefits. Focal Loss, introduced for object detection (Lin et al., 2017), down-weights the loss contribution from easy examples, focusing learning on hard cases.

\*\*Mathematical Formulation:\*\*

Standard Cross-Entropy Loss:

```

CE(p) = -log(p)

```

Focal Loss:

```

FL(p\_t) = -α(1 - p\_t)^γ · log(p\_t)