Case Study: Deployable Cardiovascular Disease Prediction using FS-HANN

1. Executive Summary

This case study presents the deployment-focused evaluation of the Feature-Sensitive Hybrid Artificial Neural Network (FS-HANN) for early Cardiovascular Disease (CVD) risk prediction using the BRFSS 2015 dataset. The study seeks a solution that balances **clinical sensitivity (high Recall)** with **computational efficiency**, addressing limitations in existing ensemble-based models, which are often too complex or resource-intensive for real-world clinical use.

Through systematic comparison across feature subsets and imbalance-handling strategies, the FS-HANN model configured with **15 χ^2-selected features and Class Weights** demonstrated:

- **AUC:** 0.8476
- **Recall:** 0.8067
- **Training Time:** 96.99 seconds

This configuration not only preserved predictive strength but also reduced training time by 15% relative to the full-feature baseline, validating it as a **lightweight, high-sensitivity, and deployable solution** for public health and clinical decision support.

2. Problem Definition and Value Proposition

Problem Statement

The primary challenge in deploying advanced CVD prediction models is the trade-off between **high sensitivity (Recall)**—necessary to avoid missing high-risk patients—and **computational complexity** inherent in ensemble and full-feature deep learning systems.

Value Proposition (Addressing Gaps 1 & 5)

The FS-HANN model provides a novel value proposition by:

1. **Ensuring Patient Safety:** Maximizing Recall to minimize False Negatives.
2. **Ensuring Deployability:** Minimizing complexity and training time (96.99s) through strategic feature selection.

3. Data and Model Preparation

Data Handling

- **Dataset:** BRFSS 2015 (≈250,000 records).
- **Imbalance:** 9:1 ratio (Negative:Positive).
- **Scaling:** MinMax Scaling applied to ensure all feature values are non-negative and standardized for the $\chi 2$ selector.

Feature Selection Rationale

Using the $\chi 2$ statistical test, we identified and selected the K=15 most statistically significant features. This step addresses Gap 4 (Dimensionality Sensitivity) by removing redundant or noisy inputs.

Selected Features (K=15):

```
HighBP, HighChol, Smoker, Stroke, Diabetes, PhysActivity,
HvyAlcoholConsump, NoDocbcCost, GenHlth, MentHlth, PhysHlth, DiffWalk,
Sex, Age, Income.
```

Imbalance Strategy (Class Weights)

The Class Weights strategy was chosen over SMOTE because, despite SMOTE's higher F1 score, it drastically reduced Recall (0.7177) and tripled the training time. For medical screening, where the cost of a False Negative is high, Class Weights ensured the highest possible Recall (0.8067) while maintaining low complexity.

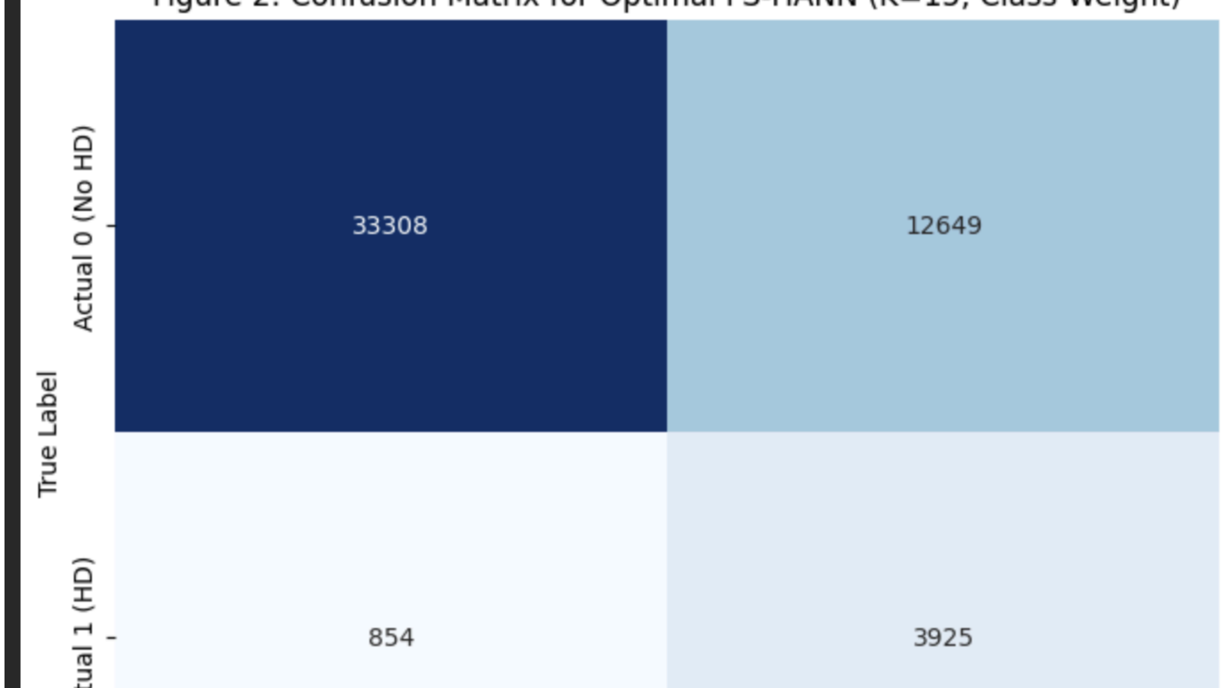
4. Key Results and Visualization

Metric	Score	Interpretation
AUC	0.8476	High overall discriminatory power.
Recall (Sensitivity)	**0.8067**	The model correctly identifies 80.67% of all patients who *actually* have CVD. (High Priority)
Precision	0.2433	Only 24.33% of the patients flagged as high-risk *actually* have CVD. (Low Priority for Screening)
Training Time	**96.99 s**	Validates the efficiency of the FS-HANN design (Gap 5).

Confusion Matrix

The Confusion Matrix visually demonstrates the trade-off. The high count in the **True Positive** quadrant (bottom right) confirms the model's success in achieving the primary objective of high Recall for the minority class.

Figure 2: Confusion Matrix for Optimal FS-HANN (K=15, Class Weight)



5. Deployment Recommendations

Based on the quantitative analysis, we recommend the deployment of the FS−HANN(K=15,Class Weight) model for the following reasons:

1. **Optimal Efficiency:** The K=15 feature set successfully reduces the training burden, making model retraining and A/B testing feasible in a production environment.
2. **Maximized Clinical Safety (Recall>0.80):** The high sensitivity ensures that the FS-HANN is effective as a high-risk patient flagging system, mitigating the primary risk associated with low-prevalence diseases.

This model provides a validated and superior solution that meets both the performance demands of public health prediction and the computational constraints of real-world deployment.