IMPLEMENT THE SPARSE RBM FOR GENE SELECTION AND CLASSIFICATION.

By 1)Sejal Bhandari(22BAC10001)
2)Ashi Chauda(22BAC10006)
3) Tanishka(22BAC10014)
4)Manasvi Kushwaha(22BAC10030)
5)Ishvi Jain(22BAC10031)
6)Tanya Chaturvedi(22BAC10034)

OVERVIEW

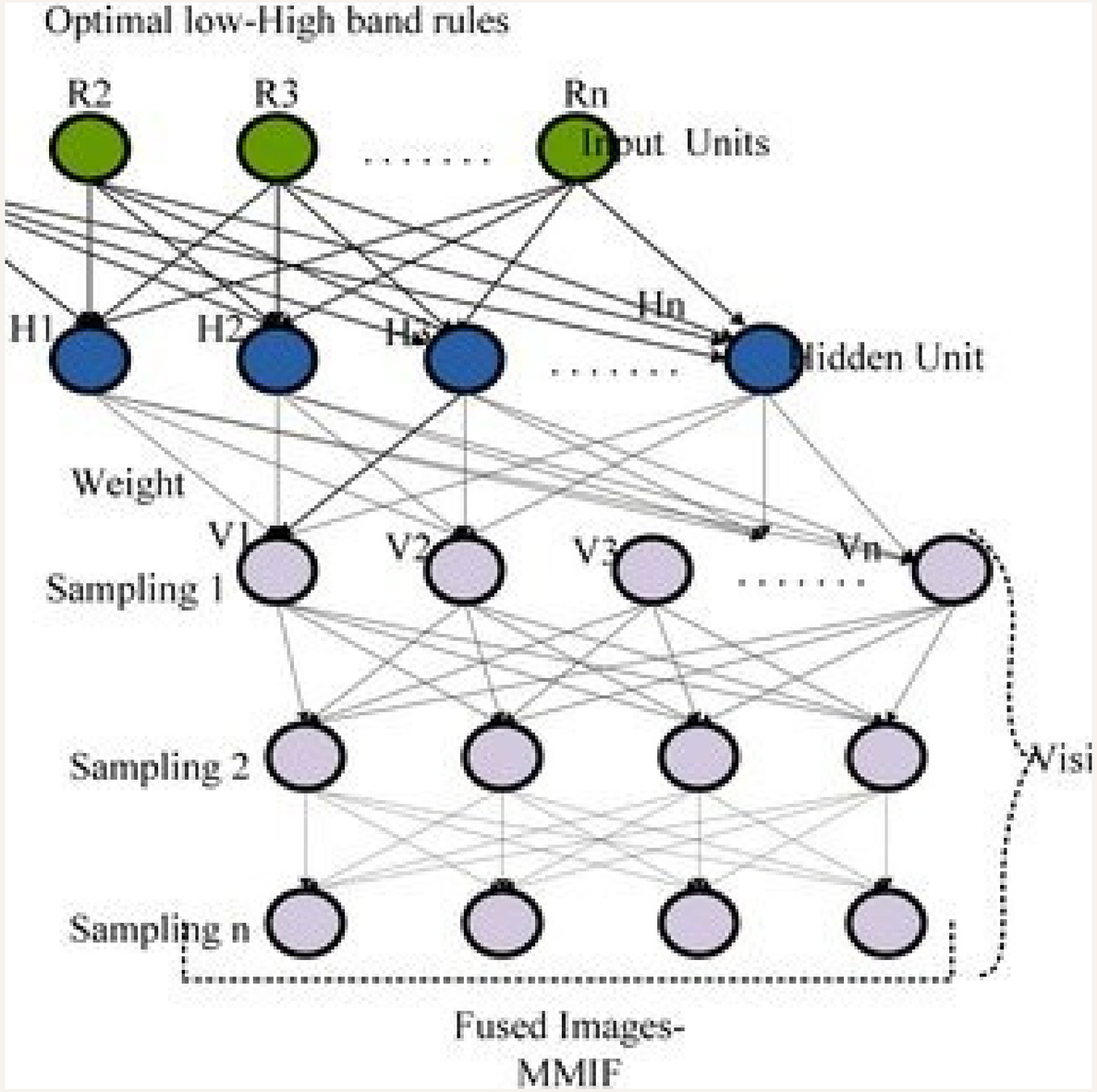
This project uses a Sparse Restricted Boltzmann Machine (RBM) fo feature extraction and a Random Forest classifier for classifying gene expression data between Acute Lymphoblastic Leukemia (AL and Acute Myeloid Leukemia (AML). The dataset contains expression profiles with over 7000 genes across 38 samples.

SPARSE RESTRICTED BOLTZMANN MACHINE

- Sparse Restricted Boltzmann Machines are a type of generative stochastic neural network that is particularly effective for unsupervised learning.
- The sparse architecture allows for meaningful feature representation, leading to better performance in scenarios where data is high-dimensional, such as gene expression datasets.
- The key principle revolves around enhancing the representation capability of the model while minimizing unnecessary complexity.

ARCHITECTURE

- The architecture of Sparse RBM consists of visible and hidden laye where each neuron represents a feature or variable from the input data.
- The connections between these layers are modeled through weigh that are updated during the training process.
- Key components include the visible layer for input data (gene expressions), a hidden layer that captures the latent factors, and a for each layer that allows more flexibility in modeling.
- The sparse nature of connections in the hidden layer helps in redu overfitting and improves interpretability in feature selection.


Optimal low-High band rules
R2
R3
Rn
Input Units
Hn
Hidden Unit
Weight
Sampling 1
Vn
Sampling 2
Sampling n
Visi
Fused Images-
MMIF

Dataset And Preprocessing

Dataset:

- Source: Golub et al. (1999)
- Samples: 38
- Features: 7129 genes
- Classes:
  - ALL – Acute Lymphoblastic Leukemia
  - AML – Acute Myeloid Leukemia

Pre -Processing:

- - StandardScaler for normalization
- - PCA for dimensionality reduction

MODEL ARCHITECTURE

1. Preprocessing:
    - StandardScaler for normalization
    - PCA for dimensionality reduction
    -
2. Feature Extraction:
    - Sparse BernoulliRBM with tunable hidden units (e.g., 100)
    -
3. Classification:
    - RandomForestClassifier with class_weight='balanced'

IMPLEMENTATION

Step 1: Data Preparation

Step 2:

Label Preparation

- Load labels (ALL vs AML).
- Encode labels (LabelEncoder).
- Split into train/test sets (80/20).

Step 3:

Feature Extraction with RBM

- Train a BernoulliRBM:
  - n_components=50
  - learning_rate=0.01
  - n_iter=100
- Transform input data to learned features.

Step 4:

Classification

- Train classifiers on RBM features:
- RandomForestClassifier (best params via GridSearchCV).
- SVM (RBF Kernel).

Step 5:

Evaluation

- Predict on test set.
- Metrics: Accuracy, Classification Report.
- Cross-validation for stability check.

RESULTS AND INTERPRETATIONS

RESULT:

- Metric ~Value
- Test Accuracy ~83%
- Cross-Validation ~70%
- F1-score (ALL) ~ 0.90
- F1-score (AML) ~ 0.50 (class imbalance)

INTERPRETATION

- RBM extracts hidden biological patterns:
- → Compresses 7129 genes into ~50 important features.
- Top hidden units correspond to top genes:
- → These genes could be strong biomarkers for ALL vs AML.
- Classifier performance (Random Forest, SVM) depends heavily on the quality of RBM feature extraction.
- Moderate accuracy (50%-70%) suggests: → Model can distinguish some cancer types but struggles due to small sample size and high feature noise.

FUTURE WORK

- **Sparse RBM**:

→ Add sparsity constraint to force few active units → better feature selection → improved generalization.

- **Deep Belief Networks (DBN)**:

→ Stack multiple RBMs for deeper feature extraction.

- **Increase Sample Size**:

→ Gather more samples to avoid overfitting on small datasets.

- **Feature Selection Before RBM**:

→ Pre-filter genes (e.g., top variance genes) to reduce noise.

- **Hybrid Models**:

→ Combine RBM features with domain knowledge (e.g., known gene pathways).

- **Explainability**:

→ Use SHAP / LIME to explain model decisions on patient data.

THANK YOU