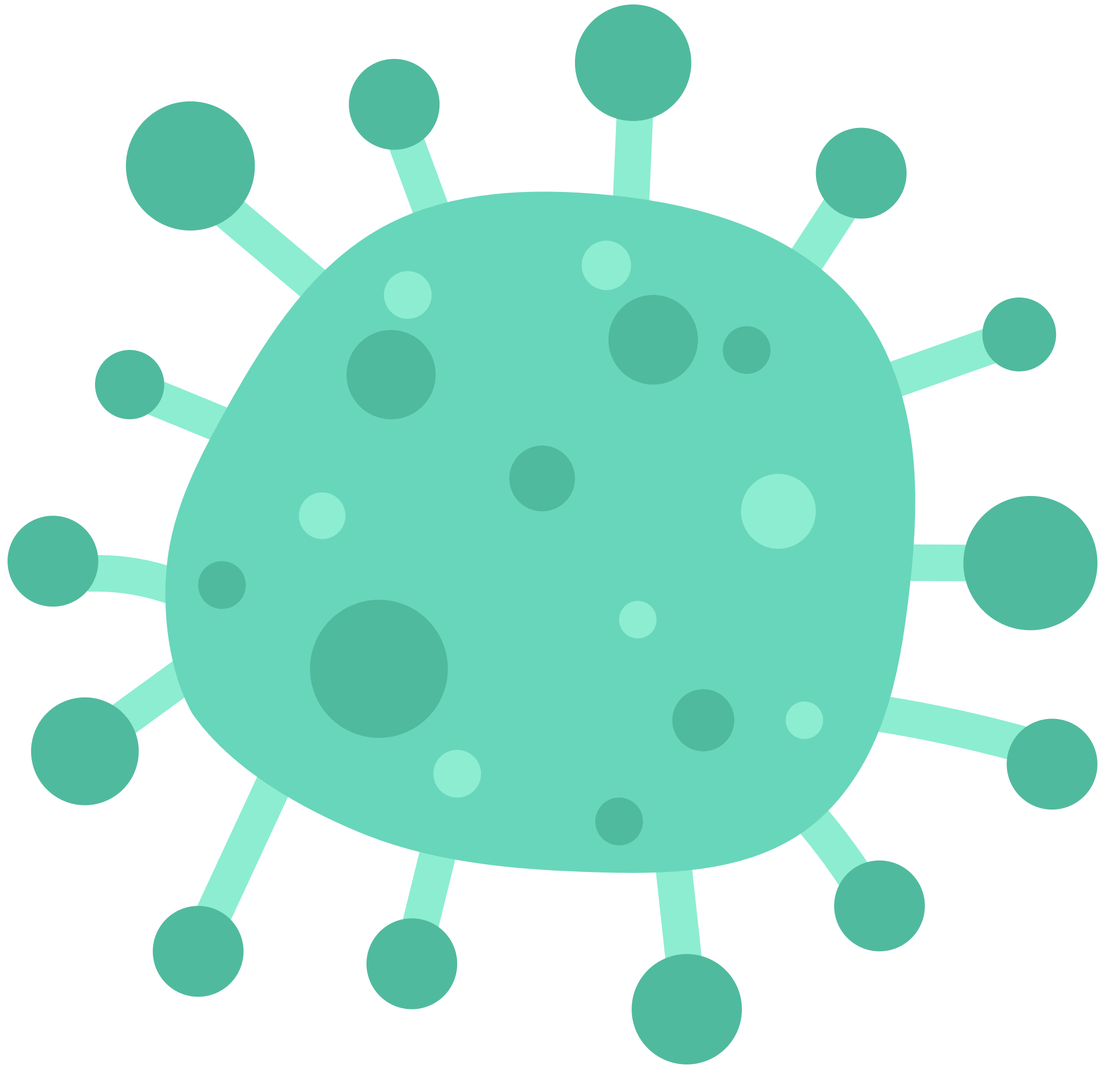


Disease Prediction with Microbial Profiles

Problem Statement

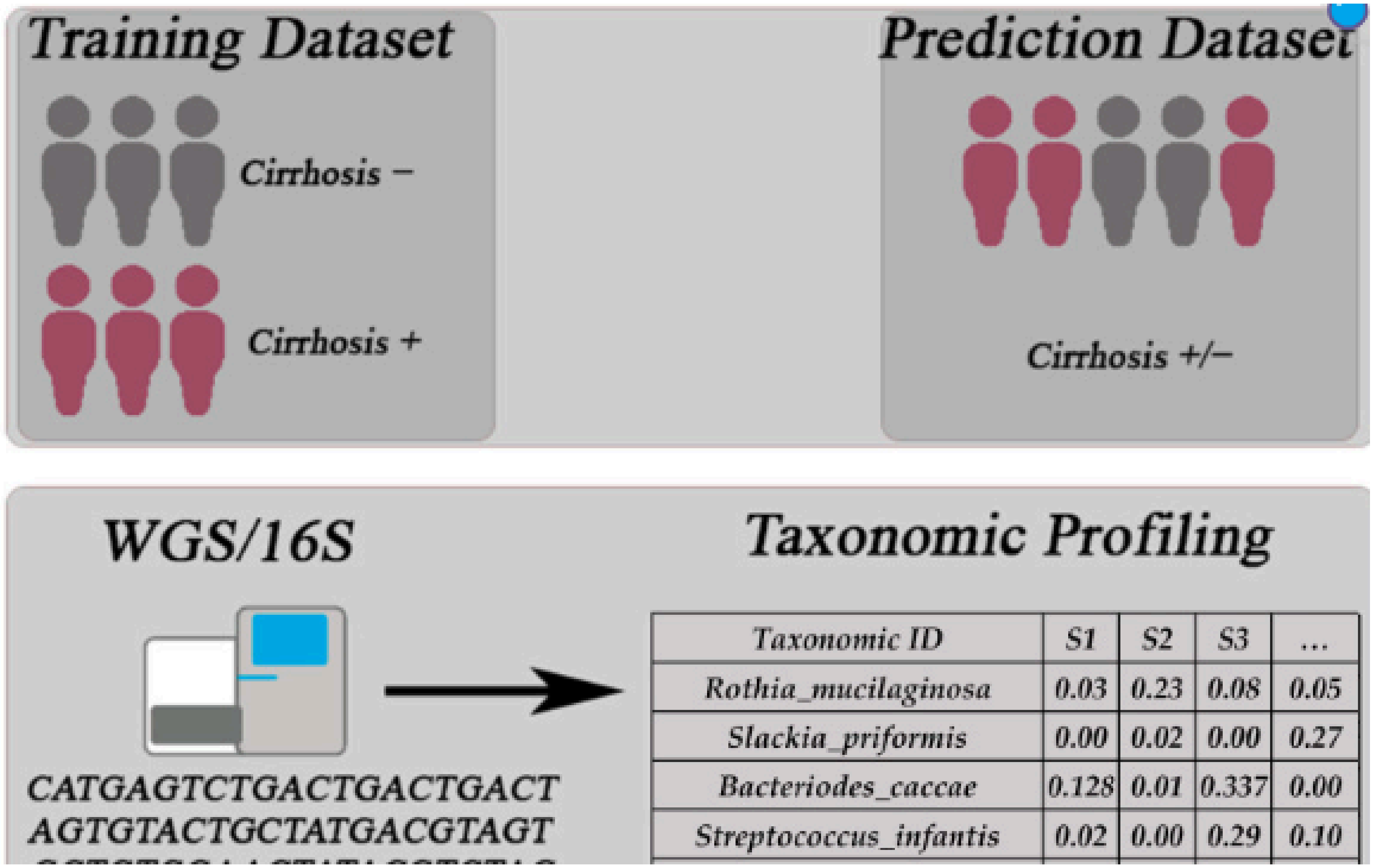
How can diseases be accurately predicted based on the microbial compositions found in patient samples, taking into account the complexity and variability of microbial data?

Dataset Overview

The dataset comprises of taxonomic profiles of the microbial communities present in each sample.

Taxonomic Profiling Process:

- Sample Collection: Bacterial samples are gathered from subjects.
- DNA Extraction: DNA is extracted from the bacterial cells.
- Sequencing: The extracted DNA nucleotides are sequenced - A, T, C, G.
- Database Comparison: The sequenced DNA is identified through reference



Taxonomic ID	S1	S2	S3	...
Rothia_mucilaginosa	0.03	0.23	0.08	0.05
Slackia_priformis	0.00	0.02	0.00	0.27
Bacteriodes_caccae	0.128	0.01	0.337	0.00
Streptococcus_infantis	0.02	0.00	0.29	0.10

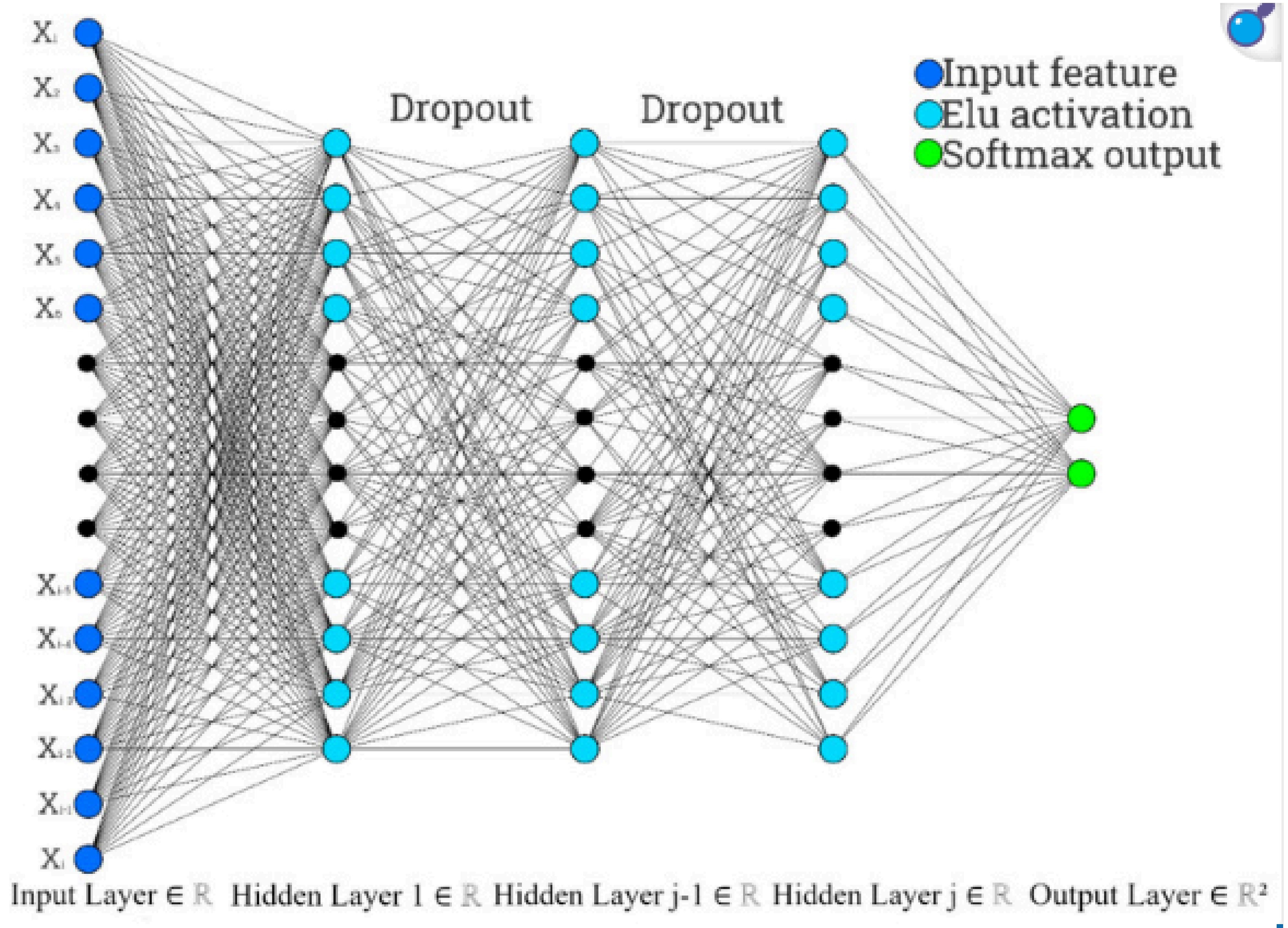
- Abundance Calculation: The relative abundance of each identified microbial species is calculated

Dataset Specification

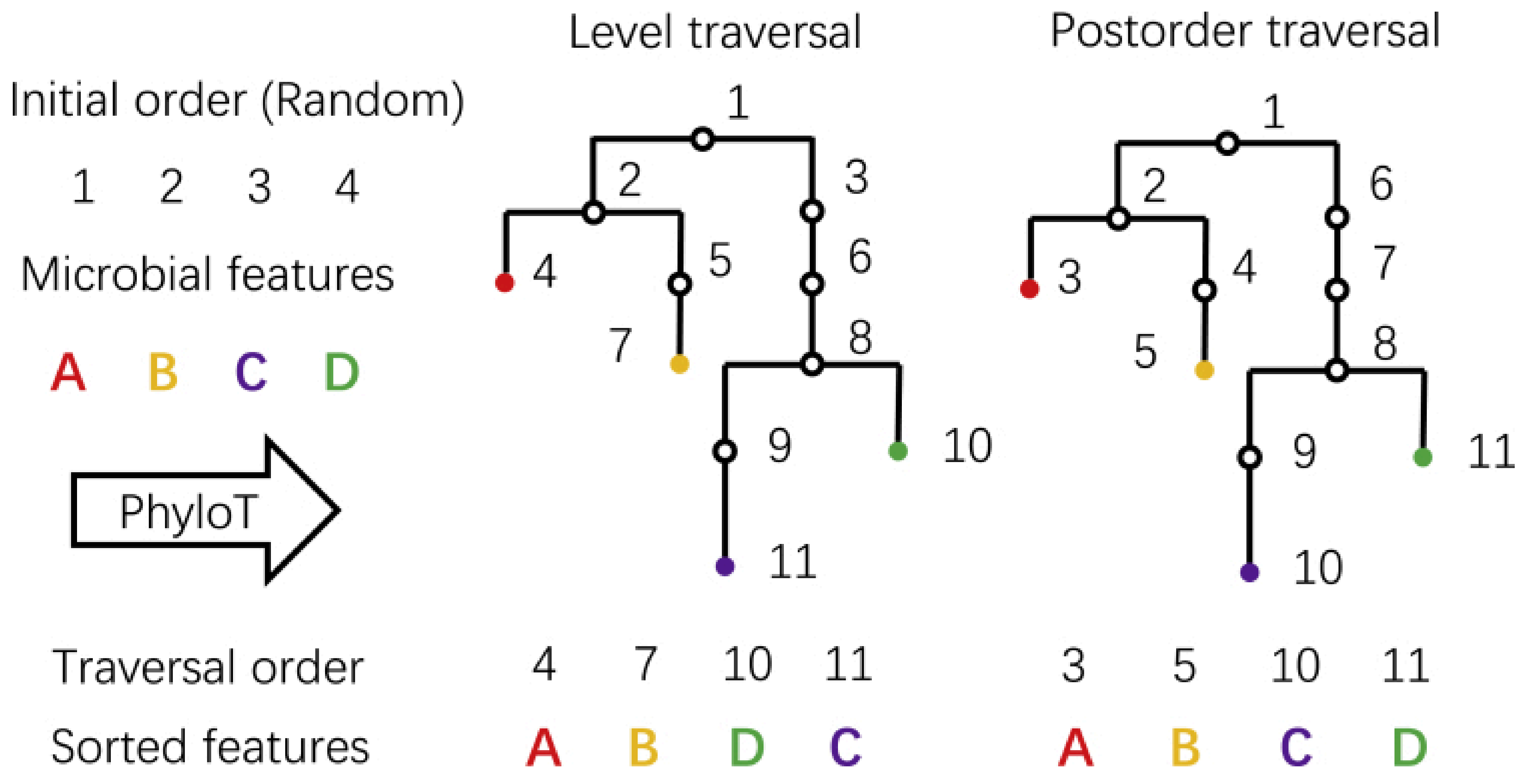
Condition	Bacterial Profiles	Samples	Train-Validation-Test Split
Type 2 Diabetes	606	440	80-10-10
Cirrhosis	3,000	243	80-10-10

Model Overview - DNN

Parameter	Value
Neuronsper Layer	60
Hidden Layers	15
Activation Function	ReLU
Dropout Layer	50%
Output Activation	Softmax
Optimizer	Adam
Loss Function	Cross-Entropy
Learning Rate	0.00025
Batch Size	50



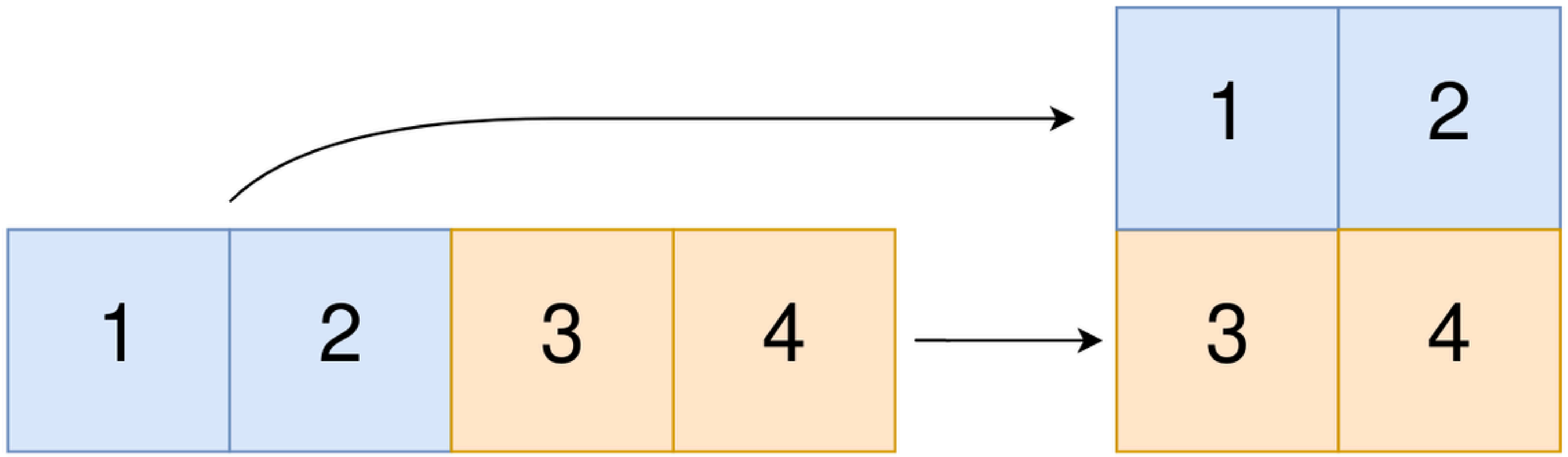
Data Preprocessing for CNNs



Phylogenetic Tree Construction through the mapping our bacteria features to NCBI dataset.

Tree Traversal (postorder & level order) to rearrange taxa with common ancestors into sequential order.

Data Preprocessing for CNNs

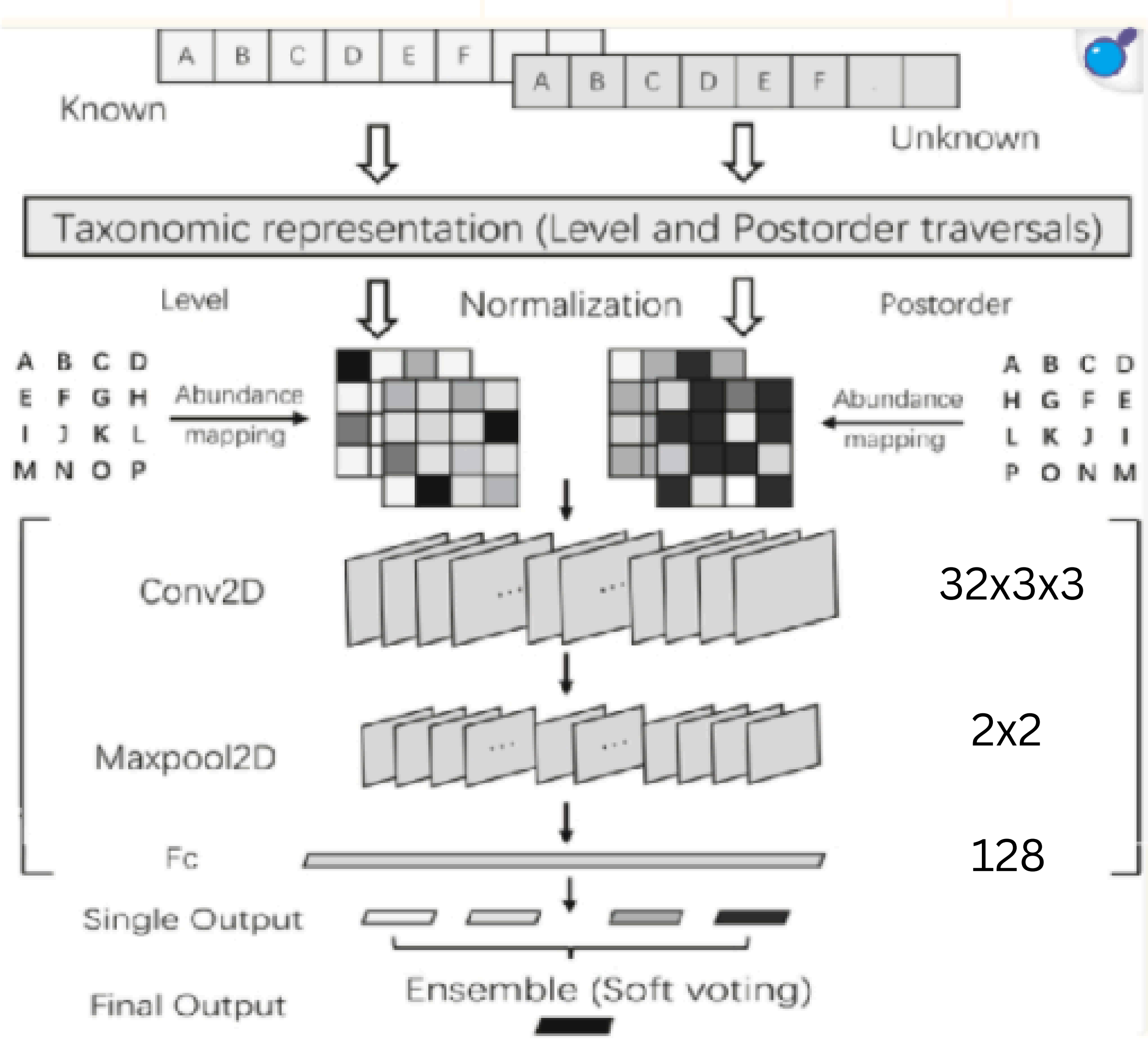


1D mapping to 2D through the formula ceil(√n) * ceil(√n) and padding the empty spaces with 0

Grayscale Coloring through the abundance of bacteria by making specific ranges

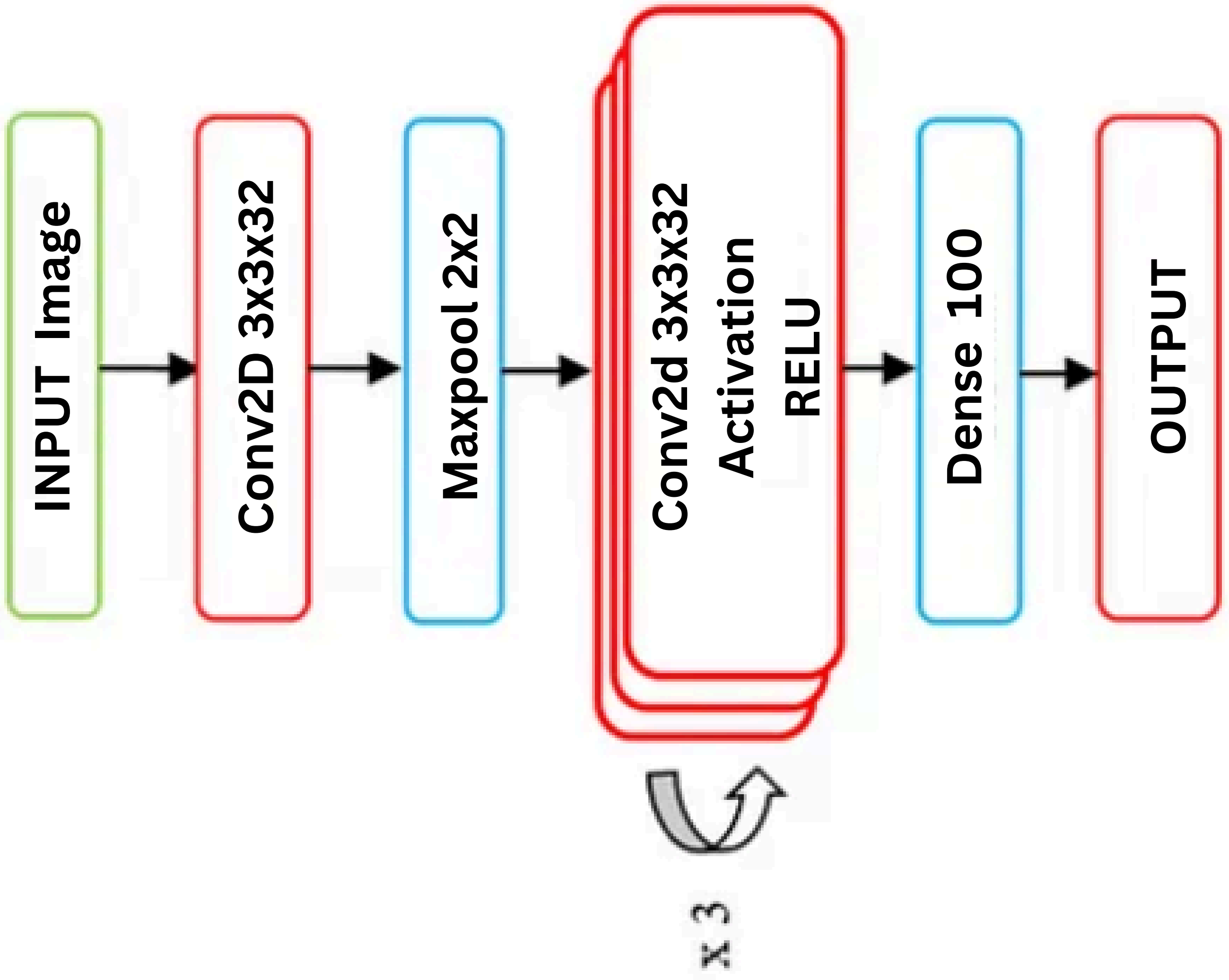
Model Overview - CNN

Parameter	Value
Activation Function	ReLU
Dropout Rate	50%
Output Activation	Softmax
Optimizer	Adam
Loss Function	BinaryCross-Entropy
Learning Rate	0.0001
Batch Size	32



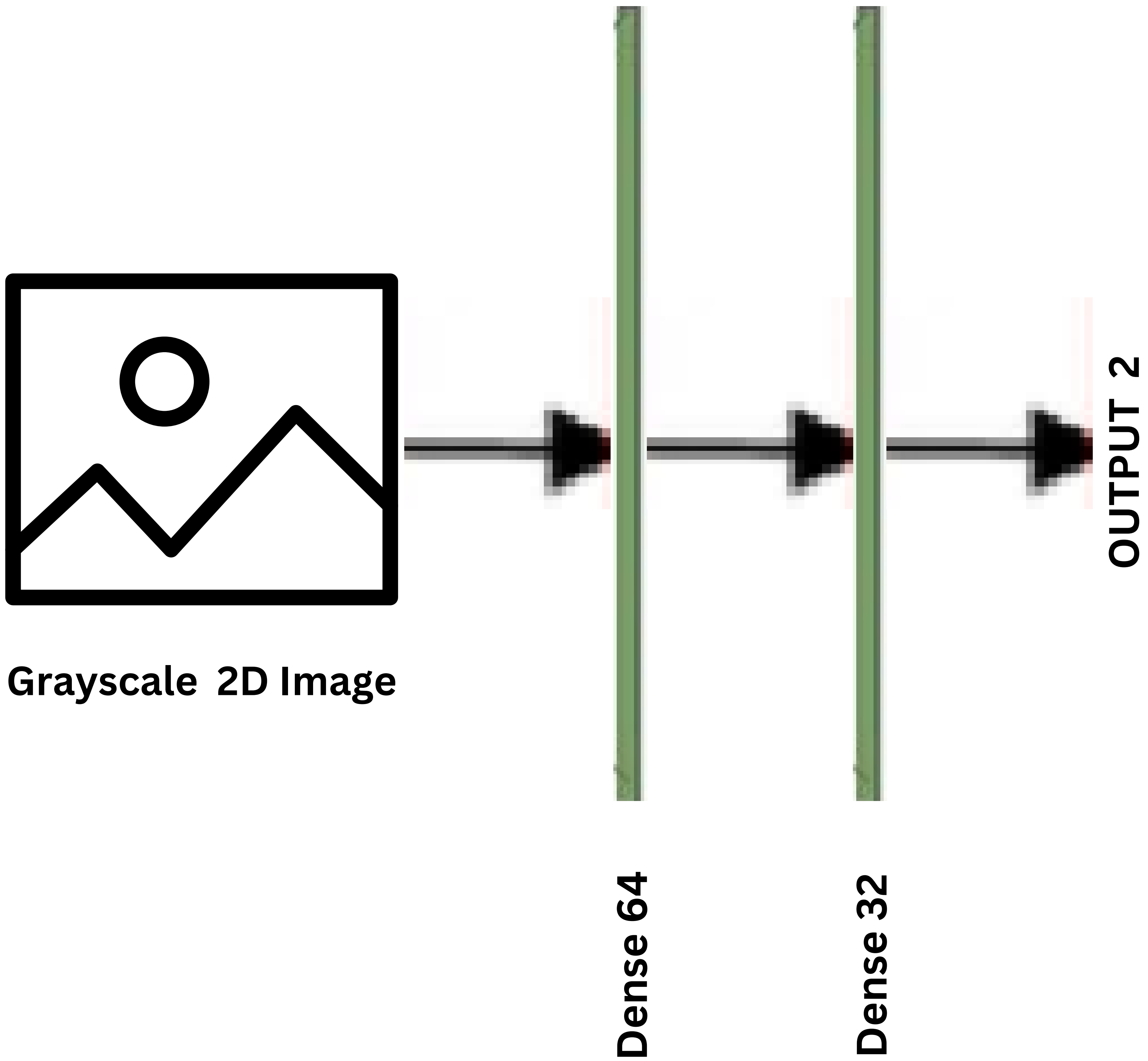
Model Overview - ResNet

Parameter	**Value**
Activation Function	ReLU
Dropout Layer	50%
Output Activation	Softmax
Optimizer	Adam
Loss Function	BinaryCross-Entropy
Learning Rate	0.0001
Batch Size	32



Model Overview - FCN

Parameter	Value
Activation Function	ReLU
Dropout Layer	50%
Output Activation	Softmax
Optimizer	Adam
Loss Function	BinaryCross-Entropy
Learning Rate	0.0001
Batch Size	32

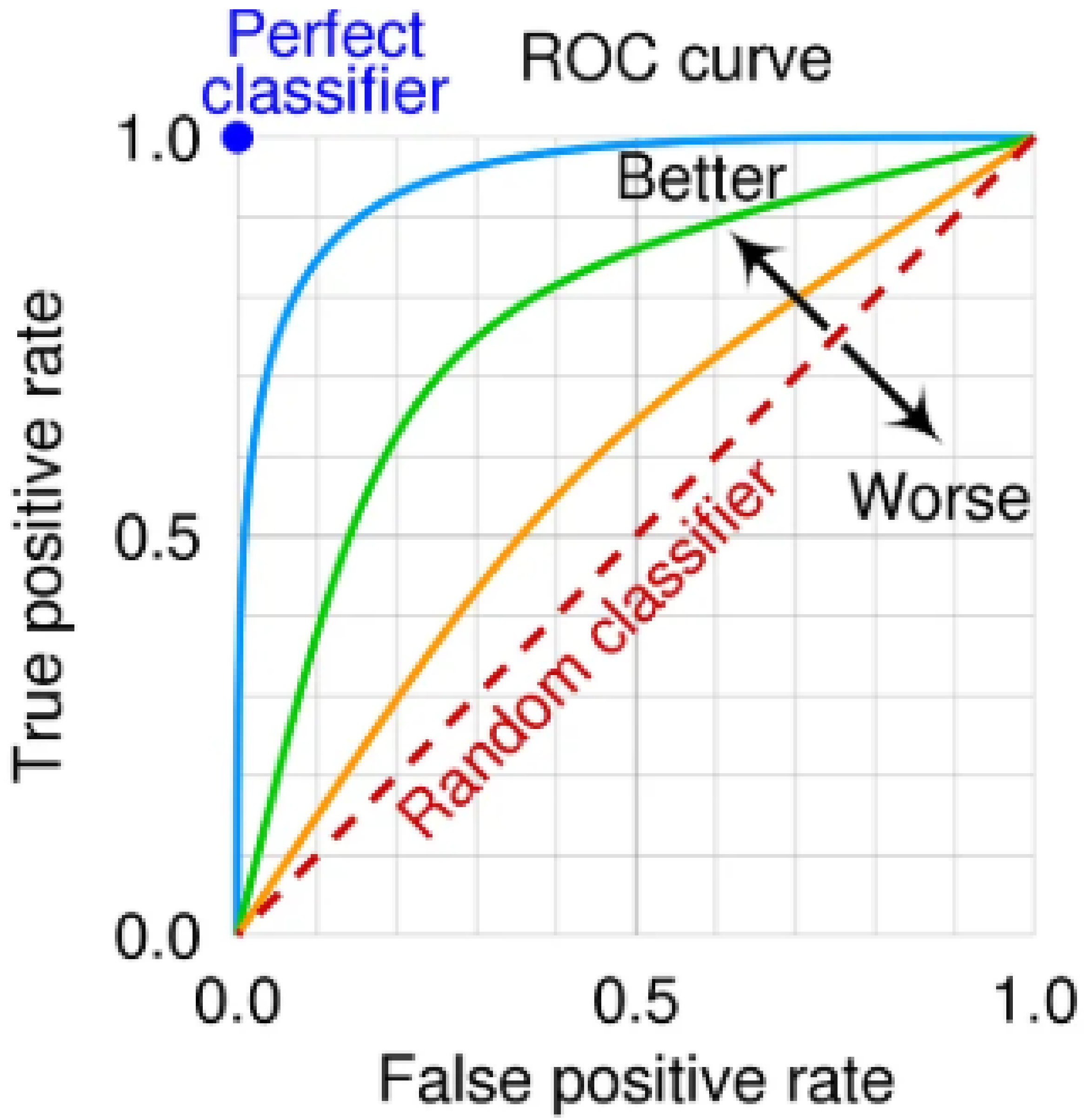


Evaluation Metrics

Accuracy: Proportion of correctly predicted instances.

AUC (Area Under the ROC Curve) measures and plots the True Positive Rate (TPR) vs. False Positive Rate (FPR) at various thresholds. It ranges from 0 to 1, with higher values indicating better performance:

- AUC = 1: Perfect classifier.
- AUC = 0.5: No discriminative ability (random guessing).
- AUC < 0.5: Worse than random guessing (predictions are reversed).



Results

Parameter ->	AUC		Accuracy (%)	
	Diabetes - T2	Cirrhosis	Diabetes - T2	Cirrhosis
KIA-DNN	0.76	0.90	70	88
KIA-CNN	0.81	0.95	72	91
KIA-ResNet	0.78	0.96	68	89
KIA-FCN	0.69	0.95	65	85
DeepForest	0.76	0.75	-	-
WRF	0.7890	0.8183	-	-
EPCNN	0.82	0.94	-	-
MegaR	-	-	67	88.5
MegaD	-	-	70	83.3
PopPhy	-	-	65	91

Results Discussion

- CNN models performed better overall for all datasets.
- Less complicated structures outperformed previously built complex models due to overfitting issues.
- Lower learning rates helped stop overfitting.
- Breaking down into small batches brought more efficiency & regularization.
- Our model overall performed better in terms of AUC, Accuracy as well as cost as it is much simpler to build and run
- Averaged predictions across models were useful

Future Work

- Combining Models to Make predictions using specific weights
- Getting data from reputable organizations to adjust our models according to the dataset.
- Curating bigger datasets by merging data and finding patterns in related diseases.
- Finding more relations through different bacterial families