A novel deep learning architectures to enable multi-scale whole genome sequence analysis for Alzheimer's disease dementia prediction

Eun Hye Lee[a], Taeho Jo[a*]

[a] Indiana Alzheimer Disease Research Center and Center for Neuroimaging, Department of Radiology and Imaging Sciences, Indiana University School of Medicine, Indianapolis, IN, USA

BACKGROUND

Alzheimer's disease (AD) dementia is the most common form of dementia. With the emergence of disease-modifying therapies for AD such as anti-amyloid monoclonal antibodies, the ability to predict disease risk before symptom onset has become increasingly important. Whole genome sequencing (WGS) data is a promising data form for early AD prediction, despite several analytical challenges. In this study, we introduce DuAL-Net, a hybrid deep learning framework designed to predict AD dementia using WGS data.

METHODS

We analyzed WGS data from 1,050 individuals obtained through the Alzheimer's Disease Neuroimaging Initiative (ADNI) and ADNI-WGS-2 with the Alzheimer's Disease Sequencing Project Follow-Up Study (ADSP-FUS1-ADNI-WGS-2). The dataset included 443 cognitively normal individuals and 607 AD dementia patients. The WGS data was generated using Illumina platforms and reads were aligned to the GRCh38 human reference genome. Standard preprocessing pipelines were applied, and single nucleotide polymorphisms (SNPs) with call rates below 95% or minor allele frequency below 1% were excluded. Genotype imputation was also conducted. For this study, we focused on a 50,000-base pair region surrounding the *APOE* gene on chromosome 19, resulting in a total of 14,011 SNPs.
DuAL-Net integrates two components: local probability modeling, which segments the genome into non-overlapping windows, and global annotation-based modeling, which annotates each SNP and reorganizes the WGS input to capture long range functional relationships (Figure 1A). Both components employ out of fold stacking with TabNet and Random Forest classifiers (Figure 1B). The final prediction is generated by combining local and global probabilities using an optimized weighting parameter alpha (Figure 1C).

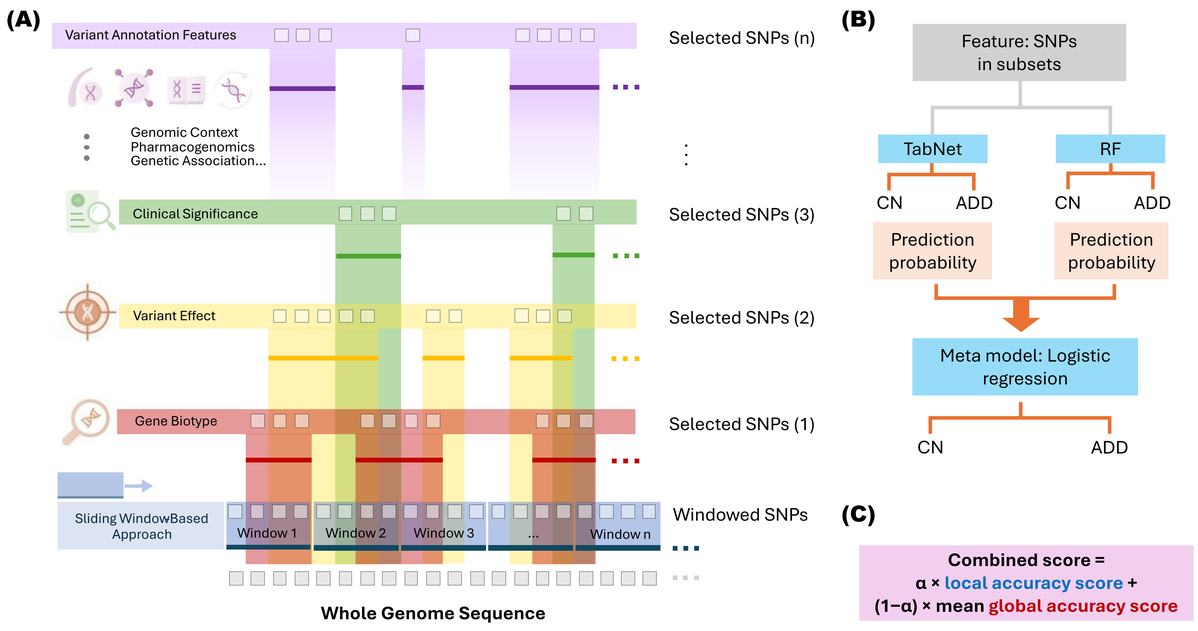

Figure 1. Annotation of genomic information to each SNP and grouping of SNPs by genomic windows or annotation categories

RESULTS

The average classification accuracy across all local SNP windows was 0.582, with accuracies ranging from 0.543 to 0.676 across the 140 windows evaluated. In the global annotation-based analysis, SNPs annotated with "pathogenic" clinical significance in ClinVar demonstrated the highest accuracy of 0.686. Variants labeled as "likely pathogenic" and "drug response" achieved accuracies of 0.676 and 0.667, respectively. Additionally, SNPs located within protein-coding genes yielded a classification accuracy of 0.657.

Using cross-validation, we identified the optimal combination weight of α = 0.8 for integrating local and global features. Under this setting, DuAL-Net achieved a cross-validated classification accuracy of 0.678.. In comparison, the global-only model (α = 0.0) reached 0.605, and the local-only model (α = 1.0) achieved 0.677.

On average across the 100, 500, and 1000 SNP subset sizes evaluated, DuAL-Net achieved an Area Under the Curve (AUC) of 0.671 using top-ranked SNPs prioritized by the model, representing 35.0% and 20.3% higher predictive performance compared to the average AUCs of bottom-ranked and randomly selected SNPs, respectively (Table 1, Figure 2). Assessment of model discriminative ability via ROC analysis across different SNP subset sizes consistently demonstrated a positive correlation between the SNPs' prioritization rank and their predictive power. The model identified SNPs with known associations to AD as top contributors to prediction, alongside potentially novel variants also ranked highly by the model.

Table 1. Comparison of accuracies by DuAL-Net with top-, bottom-ranked and randomly selected SNP subsets

SNP Subset Type	AUC (N=100)	AUC (N=500)	AUC (N=1000)	Average AUC
top-ranked SNPs	0.697	0.660	0.657	0.671
bottom-ranked SNPs	0.473	0.506	0.513	0.497
randomly selected SNPs	0.456	0.598	0.621	0.558

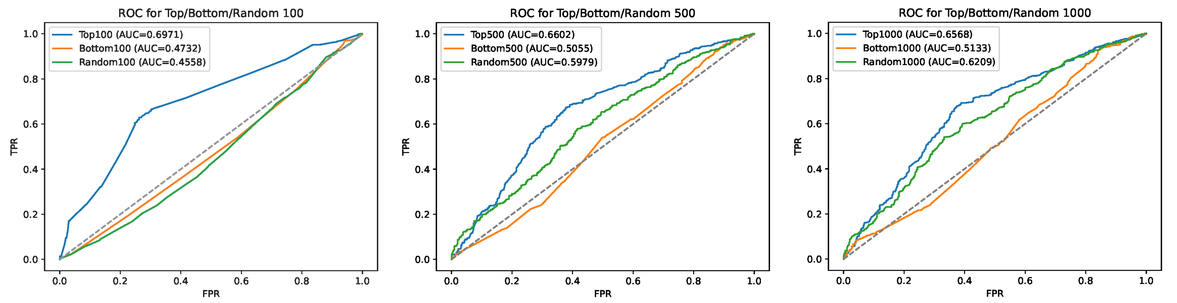

Figure 2. ROC curves comparing top-ranked, bottom-ranked and randomly selected SNP subsets identified by DuAL-Net

CONCLUSION

In conclusion, DuAL-Net presented a promising framework for AD prediction that improved predictive accuracy and enhanced biological interpretability. The framework and its web-based implementation offer an accessible platform for broader research applications.