Genomics transformer for diagnosing Parkinson's disease

1st Diego Machado Reyes
Dept. of Biomedical Engineering
Rensselaer Polytechnic Institute
Troy, New York, USA
0000-0003-2180-3251

2nd Mansu Kim
Dept. of Artificial Intelligence
Catholic University of Korea
Bucheon, Republic of Korea
0000-0002-0785-4514

3rd Hanqing Chao
Dept. of Biomedical Engineering
Rensselaer Polytechnic Institute
Troy, New York, USA
0000-0001-5973-2343

4th Juergen Hahn
Dept. of Biomedical Engineering
Rensselaer Polytechnic Institute
Troy, New York, USA
0000-0002-1078-4203

5th Li shen
Dept. of Biostatistics, Epidemiology & Informatics
University of Pennsylvania
Philadelphia, Pennsylvania, USA
0000-0002-5443-0503

6th Pingkun Yan
Dept. of Biomedical Engineering
Rensselaer Polytechnic Institute
Troy, New York, USA
0000-0002-9779-2141

***Abstract*—Parkinson's disease (PD) is the second most common neurodegenerative disease and presents a complex etiology with genomic and environmental factors and no recognized cures. Genotype data, such as single nucleotide polymorphisms (SNPs), could be used as a prodromal factor for early detection of PD. However, the polygenic nature of PD presents a challenge as the complex relationships between SNPs towards disease development are difficult to model. Traditional assessment methods such as polygenic risk scores and machine learning approaches struggle to capture the complex interactions present in the genotype data, thus limiting their discriminative capabilities in diagnosis. On the other hand, deep learning models are better suited for this task. Nevertheless, they encounter difficulties of their own such as a lack of interpretability. To overcome these limitations, in this work, a novel transformer encoder-based model is introduced to classify PD patients from healthy controls based on their genotype. This method is designed to effectively model complex global feature interactions and enable increased interpretability through the learned attention scores. The proposed framework outperformed traditional machine learning models and multilayer perceptron (MLP) baseline models. Moreover, visualization of the learned SNP-SNP associations provides not only interpretability to the model but also valuable insights into the biochemical pathways underlying PD development, which are corroborated by pathway enrichment analysis. Our results suggest novel SNP interactions to be further studied in wet lab and clinical settings.**

***Index Terms*—Parkinson's disease, Genomics, Deep learning**

I. Introduction

Parkinson's disease (PD) has a severe personal impact and economic burden on millions of people every year [1]. Coupled with its progressively debilitating nature, there are currently no recognized cures for PD [2]. While there have been major efforts to research the pathophysiologies of PD, our understanding of the disease and related disorders is still limited. Several studies agree that the combination of a person's genes and environment contributes to the risk of developing a neurodegenerative disease [3]. However, these findings are based on retrospective studies and the actual mechanisms remain to be described. Furthermore, aging is recognized as a top risk factor for most neurodegenerative diseases [4] and with an increasingly predominant aging population, neurodegenerative diseases are expected to grow in incidence and prevalence.

Parkinson's disease, like many other neurodegenerative diseases, has a complex etiology and is currently diagnosed under a differential diagnosis [5] which mainly focuses on the characteristic motor symptoms. Nevertheless, these would not appear until at least at an intermediate stage of PD. Therefore, it is key to improve the disease understanding and diagnosis ability based on prodromal factors. A very promising factor for PD diagnosis is the genotype. Nevertheless, using genotype data for the diagnosis of PD can be very challenging due to the polygenic nature of PD.

Machine learning methods have been widely employed in the genetic studies of neurodegenerative disorders [6]. Such methods can help identify disease-related genes with promising performance. However, the existing studies primarily focus on examining the main effect of individual genetic variations on the disease outcome with limited understanding of the co-occurring effects between genetic markers. Explicitly capturing the complex interactions in the genetic data contributing to the disorders is significantly under-explored. Thus, it is essential to develop new approaches to leverage the complex interactions in the genetic assessment of the disease, that, in turn, allow us to gain a deeper understanding of the biological pathways underlying Parkinson's disease. The model proposed in this work aims to bridge this gap. The complexity of the interactions between single nucleotide polymorphisms (SNPs) in a polygenic disease such as PD is a major challenge for traditional machine learning models. On the other hand,

Funding for this work was provided by NIH Training Grant (T32GM067545) supporting D.M.R. This work was also supported in part by the National Institutes of Health [R01 LM013463, P30 AG073105, U01 AG068057]; and the National Science Foundation [IIS 1837964].

neural network-based models, such as multilayer perceptron (MLP), have been shown to outperform the traditional machine learning models for PD patient classification [7]. Nevertheless, MLPs present a black-box structure limiting the interpretability of the predictive model. The community needs more advanced deep neural network models to capture these non-linear relationships in the genotype.

To bridge the gap, in this work, we propose a transformer neural network architecture for disease phenotype prediction based on the genotype data, more specifically the SNPs. The proposed transformer-based model is able to efficiently represent the data in a high-dimensional space that explicitly captures the complex interactions between the SNPs to classify PD patients from controls. The self-attention mechanism in the transformer enables to "look inside" the model, which not only increases the interpretability of the deep learning model but also provides insights to the co-occurring effects between genetic markers.

The main *methodological* contribution of this work is that the proposed model introduces transformers into the polygenic disease analysis domain. The designed transformer encoder efficiently learns and explicitly identifies the complex genomic interaction structure and increases the interpretability of the deep learning model.

In our *empirical* study, we applied the proposed model to two landmark PD biobanks: the Parkison's Progression Markers Initiative (PPMI) and the Parkinson's Disease Biomarkers Program (PDBP). The proposed model was able to achieve highly promising prediction accuracy in the PD patient classification task, outperforming traditional machine learning and deep learning methods. At the same time, our model explicitly identified a set of biologically meaningful SNP-SNP interaction patterns. These findings are highly innovative, provide valuable insights into the genetic mechanisms of PD, and can help form new hypotheses to guide subsequent molecular and clinical investigations.

II. Materials and Methods

Polygenic diseases, such as PD, present complex data patterns and feature interactions. Traditional statistical models and machine learning models struggle at capturing the high-dimensional feature interactions present in the data. Deep learning models excel at these tasks but present challenges of their own. First, the complex SNP interactions towards PD development are challenging for models to learn due to multi-factorial conditions in regulatory and coding regions in the genome related to disease development. Second, while traditional neural networks can model some of the high-dimensional feature interactions and achieve high performance, these models present limited interpretability.

In this work, to address the above challenges, we introduce the transformer to the genotype encoding domain to differentiate PD patients from controls, as it is specialized in capturing long-range semantic dependencies just like the ones present in the genome. Fig. 1 shows the overall developed framework. Using SNPs as input to the framework, the proposed transformer model for PD patient classification effectively learns and identifies the complex interactions between SNPs and provides insights into the learned SNPs relationship through the visualization of these connections. SNP data is usually encoded in an allele dosage additive representation (AA-0, AB-1, BB-2). This discrete representation is not ideal for deep learning models as it limits the level of finer details captured in the data. Thus, the first module of our framework first converts each SNP from the original additive discrete representation to a continuous variable and concurrently removes confounding effects. This is applied right after the initial quality control and PD GWAS-related SNP selection.

The second module of our framework is the proposed transformer model as shown in Fig. 1. It learns the essential relationships between SNPs towards PD phenotype prediction by constructing a meaningful high-dimensional representation of the data to classify the subjects. This module addresses the aforementioned challenges through the capabilities of transformer in capturing the long-range semantic dependencies in the genome. It helps gain deeper insight into the learned SNP relationships due to their multi-head attention mechanism. Here, the learned attention by the transformer captures the correlation between SNPs, thus reflecting the co-occurring effects between these towards PD detection. Then, the learned relationships can be used to perform downstream biological analysis, allowing for increased interpretability of the model. We then analyze and visualize the learned connectivity patterns to illustrate the interpretability of the predictions supported by known biological mechanisms. The details of our work are provided as follows.

A. SNP Representation and Filtering

In our work, SNP representation and filtering was implemented through the data munging module of the GenoML pipeline developed by [8]. Data quality control and PD-related SNP filtering are provided in Section III-A together with the dataset. The SNP representation module aims to convert the original allele-dosage discrete encoding to a continuous format and remove the confounding effects in the data. It first computes the principal components and then fits a linear regression using those components to represent each sample. The residual difference between the original sample and the regressed approximation is used as the final representation of data samples to input to the networks. The intuition behind this process is to remove the latent population substructure and experimental covariates with the residual variance representing the more generalizable and relevant data.

Specifically, let $G = \{\boldsymbol{g}_i \in \mathbb{R}^N | i = [1, \ldots, M]\}$, where M is the number of SNPs after data preprocessing and N is the number of subjects. The SNP representation module then applies principal component analysis (PCA) to project G onto its first 10 principal components. We denote the projected data with $G_P = \{\boldsymbol{g}_i^P\}_{i=1}^{10}$. Next, for the i-th SNP, the SNP representation module linearly regresses $\boldsymbol{g}_i$ with G_P: $\boldsymbol{g}_i{}' = W_r^i G_P^T + \boldsymbol{b}_r^i$, where W and b are the weight

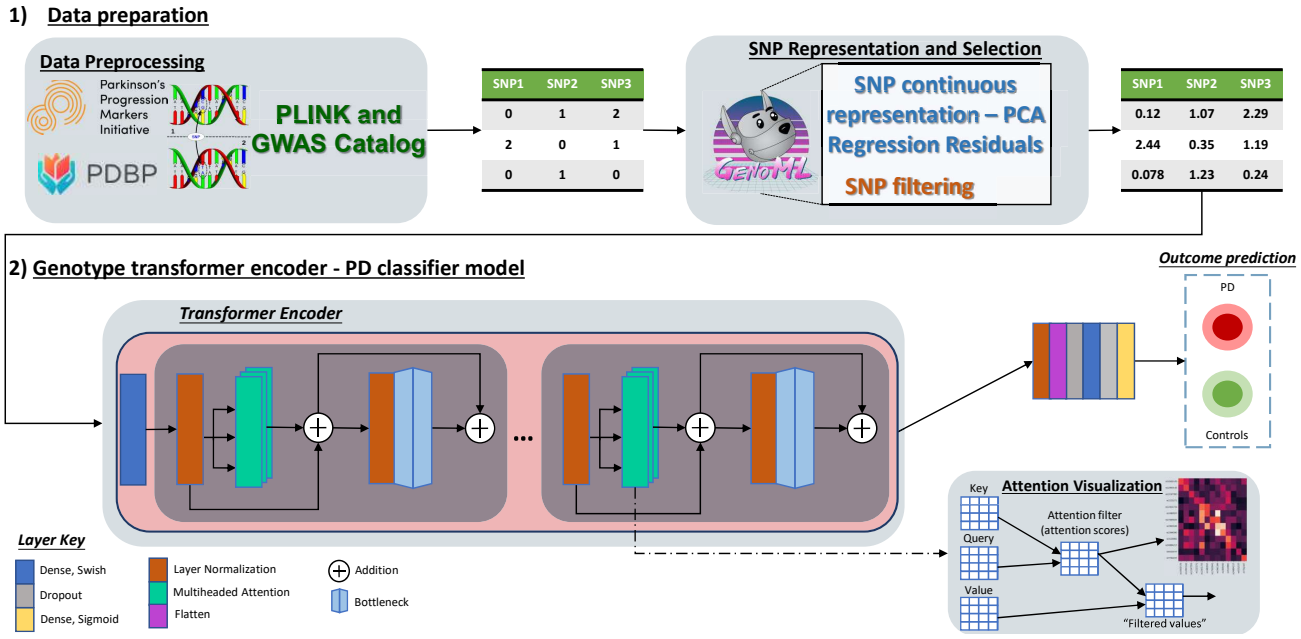

Fig. 1. Proposed framework. Module for data preparation prior to transformer model classifier, and the proposed genotype transformer encoder framework for PD patient classification.

and intercept respectively, and calculates the residual of the regression results: $\boldsymbol{r}_i = \boldsymbol{g}_i{}' - \boldsymbol{g}_i$. The final representation of the i-th SNP is computed as the normalized $\boldsymbol{r}_i$ with mean of 0 and standard deviation of 1. The SNP filtering module was implemented through the GenoML pipeline [8] to reduce the number of features to the most relevant ones using an extra tree classifier [9] to rank feature importance and select the most relevant towards PD classification.

B. Transformer Encoder Model

Our transformer consists of three modules. First, each of the pre-processed scalar SNPs is embedded into a high dimensional vector by a fully connected (FC) layer. Then, taking these vectors as input, a multi-layer transformer encoder extracts features for data representation. Finally, based on the features, an MLP with a sigmoid classifier makes PD phenotype predictions.

Let $\boldsymbol{x} = \{x_i\}_{i=1}^{m}, x_i \in \mathbb{R}$ represents the input of our transformer, where m is the number of selected SNPs. The embedding stage can be formulated as: $\boldsymbol{e_i} = W_e x_i + \boldsymbol{b}_e$, where $\boldsymbol{e}_i \in \mathbb{R}^{d_e}$ denotes the embedded d_e dimensional vectors of the i-th SNP, $W_e \in \mathbb{R}^{d_e \times 1}$ and $\boldsymbol{b}_e \in \mathbb{R}^{d_e}$ are learnable parameters shared across all SNPs.

The transformer encoder is constructed by several layers with identical components, as illustrated in Fig. 1 by the two light gray boxes inside the pink box. Each of these layers contains two sub-layers comprised in part a layer of normalization and residual connection (denoted by the addition symbol). In further detail, the first sub-layer contains a multi-head attention block, and the second sub-layer a feed-forward block (denoted as the bottleneck icon). Each head in a multi-head attention block first generates query, key, and value vectors for each SNP. Then, for each query, an output is calculated as a weighted sum of all value vectors, where the weights are computed as the similarity between the query and each key. Such an operation enables the multi-head attention block to aggregate information across all SNPs according to the query. Let $F = \{\boldsymbol{f}_i\}_{i=1}^{m}, \boldsymbol{f}_i \in \mathbb{R}^{d_{model}}$ denotes the features of SNPs input to the attention block. For the j-th attention head, the query $Q^j = \{\boldsymbol{q}_i^j\}_{i=1}^{m}$, key $K^j = \{\boldsymbol{k}_i^j\}_{i=1}^{m}$ and value $V^j = \{\boldsymbol{v}_i^j\}_{i=1}^{m}$ vectors of each SNPs is calculated by $\boldsymbol{q}_i^j = W_Q^j \boldsymbol{f}_i$, $\boldsymbol{k}_i^j = W_K^j \boldsymbol{f}_i$, and $\boldsymbol{v}_i^j = W_V^j \boldsymbol{f}_i$ respectively, where $W_Q^j, W_K^j \in \mathbb{R}^{d_k \times d_{model}}$, and $W_V^j \in \mathbb{R}^{d_v \times d_{model}}$ are learnable parameters. The output of the j-th attention head on the i-th query is computed as:

$$\text{head}_i^j(\boldsymbol{q}_i^j, K^j, V^j) = \text{softmax}\left(\frac{\boldsymbol{q}_i^{jT} K}{\sqrt{d_k}}\right) V^T. \quad (1)$$

The outputs of all attention heads are then concatenated and projected to get the final output of the multi-head attention block on the i-th token, $\text{MHA}(\boldsymbol{q}_i^j, K^j, V^j) = \text{concat}(head^1, ..., head^h) W_O$, where $W_O \in \mathbb{R}^{hd_v \times d_{model}}$ is a learnable projection matrix. Since each head has respective parameters, the multi-head attention block is able to jointly consider different types of correlations in the input feature [10]. The feed-forward block is an inverse bottleneck structure composed by two FC layers with a swish activation function [11] in between layers: The output dimension of the first FC layer d_{ff} is larger than d_{model}. The full process of the l-th layer in our transformer encoder is formulated as:

$$F_l' = \text{MA}(\text{LN}(F_{l-1})) + F_{l-1}, \quad (2)$$

$$F_l = \text{FF}(\text{LN}(F_l')) + F_l', \quad (3)$$

where $\mathrm{LN}(\cdot)$ is the layer normalization [12].

The output of the transformer encoder is a matrix in dimension of $d_{model} \times m$. It is flattened into a vector and fed to an MLP with two FC layers to produce the final prediction.

The framework was trained using a focal loss [13] and AdamW [14] optimizer. Focal loss was used for this framework due to its great capability of dealing with imbalanced datasets using weighting parameter α_t, and its capacity to focus on hard negative samples with the modulating factor $(1-p_t)^\gamma$ and focusing parameter γ. The focal loss is defined as

$$FL(p_t) = -\alpha_t(1-p_t)^\gamma log(p_t), \tag{4}$$

where p_t is the predicted probability of the ground truth class.

III. Results

A. Datasets

Two datasets were used to train and evaluate the proposed model and baselines, these are Parkison's Progression Markers Initiative (PPMI) and Parkinson's Disease Biomarkers Program (PDBP). PPMI is a well-established consortium that has collected de-identified clinical, imaging, 'omics, genetic, sensor, and biomarker data from patients with onset Parkinson's disease, at the prodromal stage, and healthy controls. Genotype data obtained from PPMI corresponded to whole-genome sequencing from whole-blood extracted DNA samples. PDPB is a study sponsored by the National Institute of Neurological Disorders and Stroke (NINDS) containing a collection of 'omics studies with the goal of accelerating the discovery of promising new diagnostic and progression biomarkers for PD. According to PDBP documentation, SNP genotyping data was obtained through the Illumina NeuroX array including exonic and additional custom variants designed for neurological disease studies.

Both datasets were made available on their corresponding websites after standard processing pipelines following current best practices. Details on the cohorts demographic distributions can be seen in Table I. It is important to notice the strong imbalance in the PPMI dataset as the ratio of PD to healthy controls (HC) participants is close to 2:1. Moreover, the age distribution between PD and HC is considerably similar, while gender presents a higher proportion of males than females. Nevertheless, no X or Y chromosome SNPs were used in the final dataset. PDBP presents a more balanced distribution of PD vs HC subjects, with a slightly higher reported age in HC subjects. Moreover, in terms of gender, the PD subjects have close to double the number of male than female subjects, while for the HC distribution the opposite case is observed.

QC on genotype data was perform current best practices for PPMI as described in [7]. SNP representation and filtering were implemented using the GenoML [8] data munging pipeline as described in Section II-A. For the SNP representation 10 principal components (PCs) were used for the PPMI dataset. The resulting dataset contained genotype data for 510 subjects and 61 SNPs. For the PDBP dataset, 1154 subjects had genotype (269,476 variants) and phenotype data available for processing. The same SNP filtering and QC pipeline from PPMI was applied to the PDBP dataset, with the only difference being it used 2 PCs. Only 2 PCs were used for PDBP as this captured an equivalent proportion of explained variance as 10 PCs in PPMI. After this process, the PDBP dataset contained 1068 subjects and 58 SNPs. Finally, the SNP overlap between both datasets was found to be 13 SNPs. The overlapping SNPs were chosen for common ground comparison across models; thus, blocking the confounding variable to have the same features used in each experimental setting. Therefore, the final datasets used as input to the models consisted of 510 subjects and 13 SNPs for PPMI, and 1068 subjects and 13 SNPs for PDBP.

TABLE I
SUBJECT DATA DISTRIBUTION

		PD	**HC**
PPMI	Participants	349	161
	Age	61.50 ± 9.56	61.27 ± 10.7
	Gender	M:227 F:122	M:104 F:57
PDBP	Participants	574	496
	Age	65.64 ± 11.9	69.97 ± 12.0
	Gender	M:379 F:195	M:190 F:306

B. Evaluation Strategy

The proposed transformer encoder model was compared against several well-established traditional machine learning models - random forest, support vector machine (SVM) with radial-basis function (RBF) kernel and logistic regression (LR), as well as a multi-layer perceptron (MLP) and long short-term memory (LSTM) for the deep learning models. The machine learning models, namely random forest rbf-SVM and LR, were implemented using sci-Kit learn python implementations and tuned using an exhaustive grid-search using the sklearn GridSearchCV API.

On the other hand, deep learning models, namely MLP, LSTM and Transformer, were implemented using Tensorflow-Keras API and tuned manually as the hyperparameter search space was too large for an exhaustive cross-validation grid-search approach. In short the manual tuning consisted of progressively choosing the best performing hyperparameters by modifying one category at a time. First, learning rate and loss parameters were tuned, then number of layers and units, to finally progressing to lower impact hyperparameters such as the dropout rate. Hyperparameters for the proposed transformer-based model and baseline models were determined using a 10-fold cross-validation approach applied to the training portion of an 80/20 train-test split. The described hyperparameter tuning process allowed for an unbiased tuning process to find near to optimal configurations for deep learning models and the optimal settings for all machine learning models. The area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC) were calculated for all evaluated models using the Sci-Kit learn. It is noteworthy that all machine learning models

were trained with a balanced class weight parameter to mimic the functionality of the focal loss on the proposed model. On the other hand, all deep learning models were trained using focal loss for fair comparison amongst them. Hyperparameter configurations for all models, and further details on the manual tuning process will be made available together with the code after acceptance.

Due to the small sample sizes, classification results vary considerably depending on the samples used for testing. In order to alleviate this confounding variable, the proposed transformer-based model and baseline models were evaluated using a 10-fold stratified cross-validation framework applied to each complete dataset to ensure accurate results, with special focus to the PPMI dataset due to its small size and imbalance. It is noteworthy that, a limitation of this work is the presence of so called data leakage as there is an overlap present between the samples used for the hyperparameter search and the 10-fold cross validation evaluation framework. Nevertheless, the impact of the confounding variability in classification results due to the testing samples used at different partitions is considerably higher. Moreover, it is noteworthy that the random states of the data splits are different at the tuning and evaluation stages; in other words, the grouping of samples vary between the tuning and evaluation stages. Therefore, a certain degree of stochasticity is introduced for the hyperparameter settings at the evaluation stage. Similarly, as the process is the same for all models, there is no unfair advantage introduced for any model. Random states for the data splitting were set to the same value across models to ensure fair comparison by training/testing on the same data splits, with different random states between hyperparameter search and the evaluation stages. The AUROC and AUPRC at each test partition were calculated for each one of the 10-fold experimental units, then the mean and standard deviations were reported as the final results. Significance testing comparing the proposed transformer and baseline models was performed using paired samples Wilcoxon test through the SciPy Python library.

C. PD Prediction Results

The classification results from the proposed networks can be seen in Table II. As shown in the table, the PDBP dataset showed to be considerably more challenging for all the models, this was surprising as both machine learning and deep learning models tend to perform better with larger datasets. However, the lower performances could be due to the different technologies used to obtain the genotype data.

In terms of the model comparisons, the proposed transformer-based model significantly outperformed all the baseline models (AUROC and AUPRC) in both PPMI and PDBP experiments. The propsed transformer model achieves higher classification results due to its design, with special focus to the self-attention, to efficiently capture the complex interactions between the SNPs towards PD development. The random forest model performed second best in all experiments, this is expected as ensemble models are very effective at classification tasks using tabular data as these can find optimal

TABLE II
10-FOLD CROSS-VALIDATION PERFORMANCE OF THE PROPOSED MODEL AND BASELINES. SIGNIFICANT IMPROVEMENT WAS FOUND USING THE TRANSFORMER MODEL COMPARED TO THE ALL BASELINE MODELS FOR THE PPMI AND PDPB DATASETS. VALUES WITH SIGNIFICANT DIFFERENCE ($\alpha = 0.05$) DENOTED WITH '*' AND '**' FOR $\alpha = 0.005$

Dataset	Model	Mean AUROC $\pm$ SD	Mean AUPRC $\pm$ SD
PPMI	RF	0.656 $\pm$ 0.095 *	0.797 $\pm$ 0.073 *
	SVM	0.595 $\pm$ 0.063 **	0.769 $\pm$ 0.073 **
	LR	0.588 $\pm$ 0.062 **	0.764 $\pm$ 0.060 **
	MLP	0.605 $\pm$ 0.066 *	0.774 $\pm$ 0.065 **
	LSTM	0.568 $\pm$ 0.081 **	0.737 $\pm$ 0.068 **
	Transformer	**0.708 $\pm$ 0.106**	**0.835 $\pm$ 0.078**
PDBP	RF	0.538 $\pm$ 0.043 *	0.566 $\pm$ 0.044 *
	SVM	0.505 $\pm$ 0.033 *	0.557 $\pm$ 0.045 **
	LR	0.468 $\pm$ 0.052 **	0.525 $\pm$ 0.045 **
	MLP	0.480 $\pm$ 0.040 **	0.524 $\pm$ 0.047 **
	LSTM	0.509 $\pm$ 0.042 *	0.542 $\pm$ 0.039 *
	Transformer	**0.581 $\pm$ 0.048**	**0.611 $\pm$ 0.030**

combinations of input features for the task at hand. The remaining models had varied performances depending on the dataset, as the MLP and SVM achieved third and fourth best results on the PPMI dataset respectively. MLP could capture some of the feature interactions, but not as efficiently as the more complex models, such as the proposed transformer-based model. Note that, the proposed transformer-based model uses MLP to perform the outcome prediction based on the learned representations. With the transformer encoder, it is able to capture the complex SNP interactions that will allow for the higher performance by the proposed model.

On the machine learning models, the SVM uses an RBF kernel to find complex high-dimensional boundaries that allows it to perform well at classification tasks with complex interactions; however, as seen in the results, it is not as efficient as deep learning models such as the proposed transformer and MLP models. Finally, the LSTM model is a natural precursor to the transformer model to aggregate information from the input features and it is able to achieve the third highest performance in the PDBP dataset.

D. Interpretability of Predictions

In addition to the PD patients and healthy controls classification, this section presents a deeper insight to increase the interpretability of the models. A key advantage of transformer models is the ability to analyze the self-attention scores produced from the key-query matrix multiplication. These attention scores provide a numeric interpretation of the relationship between features. In the case of the transformer model in this work, the attention scores are used to describe the learned SNP-SNP relationships towards the patient classification task. In order to provide a clear visualization of the learned relationships, chord plots were drawn using the circlize R library [15]. The top transformer-based models (i.e. highest AUROC in the test set) from the 10-fold cross-validation evaluation were visualized from each dataset. The learned attention scores were averaged across all subjects in

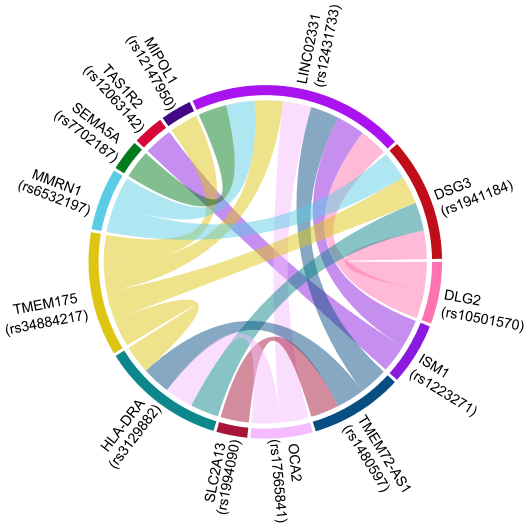

Fig. 2. Transformer learned SNP interactions on the PPMI dataset.

the corresponding test set, resulting in a mean attention matrix per head. As each head learns a different set of attention relationships, max pooling then was applied in the channel dimension of the mean attention matrices, i.e. across the heads, to summarize the most relevant learned connections for the model.

For downstream analysis of the learned SNPs relationships, the enrichment analysis tool, Enrichr, was employed to identify the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways enriched by our genetic findings. The first gene set was determined from the top three performing models trained with the PPMI dataset, from which the highest performing model can be seen in Fig. 2. In this gene set, two enriched KEGG pathways were observed (p ¡ 0.05). First, the "Carbohydrate digestion and absorption" pathway was enriched with the smallest raw p-value of 0.019 and odds ratio of 61.49. Second, the "taste induction" pathway was found to be enriched with a raw p-value of 0.034 and odds ratio of 33.46. Nevertheless, it is noteworthy that the genes corresponding to the SNPs with strongest learned relationships had no previously associated enriched KEGG pathways, namely TMEM72-AS1, SLC2A13 and MIPOL1. Further investigation to evaluate the potential of these genes as PD biomarkers is needed.

For the PDBP gene set, the "Hippo signaling pathway" was found to be enriched with p-value of 0.032 and odds ratio of 40.81. Similarly, the "tight junction" had p-value of 0.033 and odds ratio of 39.34. These two pathways were also found to be enriched in the first gene set ranking the 3rd and 4th. While the hippo signaling pathway has traditionally been associated with cancer, recent studies have shifted their attention towards this pathway's connection with neurodegenerative diseases [16]. Moreover, Recently, dysfunction in the tight junctions and their interaction with microbiota in the intestinal barrier have linked with gut dysbiosis in PD [17]. Recent studies in this field have focused on the gut-to-brain PD approach [18]. Our model found relevant connections between SNPs associated to gut-related pathways such as the carbohydrate digestion and absorption tight junction. A key area of further research would look into the connections between these pathways and elucidate on the putative genomic biomarkers.

Moreover, the individual main effects were analyzed for the proposed transformer-based model and the best performing baseline model, namely random forest. For the former an Out of Bag Feature Importance approach is taken to evaluate the impact on the performance of the model by removing one feature at a time. For the latter, the sklearn built-in feature importance, which implements gini importance, is used to rank the SNPs relevance towards the classification task. Both approaches are applied on the models trained on the highest performing folds of the 10-fold cross validation evaluation framework. For random forest, the top three SNPs in PPMI corresponded to the genes TMEM175, MIPOL1, MMRN1; while for PDBP matched LINC02331, DSG3, TMEM175. On the other, hand for the proposed transformer-based model the top three in PPMI were OCA2, DLG2 and TMEM175, while for the PDBP dataset were found to be MMRN1, SLC2A13 and TMEM175. The shared top feature across both models and both datasets, namely the Transmembrane protein 175 (TMEM175) gene, has been previously associated with PD pathogensis through a critical role in lysosomal and mitochondrial function, as neurons with TMEM175 deficiency have shown increased phosphorylated and detergent-insoluble α-synuclein deposits [19].

IV. Discussion and Conclusion

The proposed transformer model for genotype encoding and PD patient classification outperformed traditional machine learning and deep learning baseline models. Deep learning methods have been on the edge of clinical analysis due to their black box implementation. However, novel methods such as the transformer model presented in this work provide a behind-the-scenes of the deep learning model. Thus, it allows for increased interpretability of the underlying feature associations towards patient classification. The proposed transformer model learned the key relationships between the SNPs to produce a high-dimensional representation of each genotype profile to then classify it as PD or HC. The visualization from the transformer-based model attention scores showed key connections between SNPs increasing the interpretability of the model predictions in conjunction with known mechanisms. Similarly, feature relevance scores obtained from the random forest provided complementary insight towards the key SNPs that lead towards PD according to the predictive models.

While the proposed framework achieved the best performance, there are some exciting research areas to further probe with challenges to solve. A limitation of this study is the use of a small subset of PD-related SNPs. The SNP filtering process uses an extra-trees classifier to rank SNP importance in the PPMI dataset. While it could be argued that data leakage was present due to the SNP ranking process on the full dataset, the goal of this study is not to identify PD-related SNPs rather than developing predictive models that can capture the

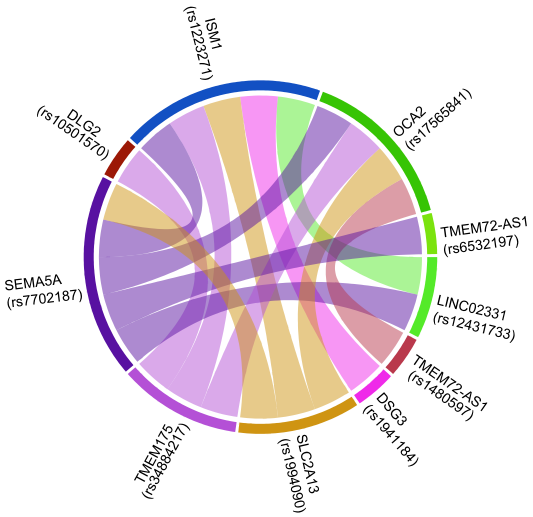

Fig. 3. Transformer learned SNP interactions on the PDBP dataset.

complex relationships in the selected SNPs. SNP filtering and SNP identification at large scale unprocessed genomic data is an exciting area of opportunity that could be integrated with predictive models, such as the one introduced in this work for sequencing-to-diagnosis pipelines.

Another limitation of this work is the small sample size. An essential challenge for biomedical data is the limited sample size, as it restrict the generalization capabilities of the deep learning networks. Current cohorts are continuously recruiting more subjects, this will aid to address the small sample sizes for training the networks. Likewise, novel training processes, such as pretraining and domain adaptation methods, could alleviate the limited sample size challenge. Modules of the network could be pretrained on larger non-PD genotype datasets for an alternative classification task, such as for ancestry prediction, and then fine-tuned towards the final outcome prediction with the specialized dataset (PPMI). This approach would model the building blocks for complex interactions in the genotype and then focus the network only on the key connections for PD.

Another exciting area of opportunity in the field is the inclusion of endophe post-transcriptional modification data that would provide the missing link between the genotype and phenotypical expression of PD. Incorporating other modalities will increase the network's ability to differentiate PD patients from controls and improve the description of the underlying mechanisms leading to PD. For example, imaging biomarkers would be another key addition to the input data. Imaging biomarkers have succeeded at differentiating PD patients from controls in previous works [6]. Lewy bodies and other imaging traits are often indicators of PD. In future work, we will integrate imaging biomarkers to further improve the proposed framework performance.

Acknowledgements

Data used in the preparation of this article were obtained from the Parkinson's Progression Markers Initiative (PPMI) database (www.ppmi-info.org/access-data-specimens/download-data). For up-to-date information on the study, visit ppmi-info.org. PPMI – a public-private partnership – is funded by the Michael J. Fox Foundation for Parkinson's Research and funding partners, including [list the full names of all of the PPMI funding partners found at www.ppmi-info.org/about-ppmi/who-we-are/study-sponsors]. Data and biospecimens used in preparation of this manuscript were obtained from the Parkinson's Disease Biomarkers Program (PDBP) Consortium, supported by the National Institute of Neurological Disorders and Stroke at the National Institutes of Health.

References

[1] W. Zahra *et al.*, "The Global Economic Impact of Neurodegenerative Diseases: Opportunities and Challenges," in *Bioeconomy for Sustainable Development*, C. Keswani, Ed. Singapore: Springer Singapore, 2020, pp. 333–345.

[2] F. Durães, M. Pinto, and E. Sousa, "Old Drugs as New Treatments for Neurodegenerative Diseases," *Pharmaceuticals*, vol. 11, no. 2, p. 44, May 2018.

[3] A. R. Dunn, K. M. O'Connell, and C. C. Kaczorowski, "Gene-by-environment interactions in Alzheimer's disease and Parkinson's disease," *Neuroscience & Biobehavioral Reviews*, vol. 103, pp. 73–80, Aug. 2019.

[4] Y. Hou *et al.*, "Ageing as a risk factor for neurodegenerative disease," *Nature Reviews Neurology*, vol. 15, no. 10, pp. 565–581, Oct. 2019.

[5] M. Papadakis, S. McPhee, and M. Rabow, *CURRENT Medical Diagnosis and Treatment 2021*, 60th ed. New York: McGraw-Hill Medical, 2020.

[6] L. Shen and P. M. Thompson, "Brain imaging genomics: Integrated analysis and machine learning," *Proc IEEE Inst Electr Electron Eng*, vol. 108, no. 1, pp. 125–162, 2020.

[7] M. B. Makarious *et al.*, "Multi-modality machine learning predicting parkinson's disease," *npj Parkinsons Dis.*, vol. 8, no. 1, pp. 1–13, number: 1 Publisher: Nature Publishing Group.

[8] M. Makarious *et al.*, "GenoML: Automated Machine Learning for Genomics," *arXiv:2103.03221 [cs, q-bio]*, Mar 2021, arXiv: 2103.03221.

[9] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, no. 1, pp. 3–42, Apr. 2006.

[10] A. Vaswani *et al.*, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.

[11] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for Activation Functions," *arXiv:1710.05941 [cs]*, Oct. 2017, arXiv: 1710.05941.

[12] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer Normalization," *arXiv:1607.06450 [cs, stat]*, Jul. 2016, arXiv: 1607.06450.

[13] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," *arXiv:1708.02002 [cs]*, Feb. 2018, arXiv: 1708.02002.

[14] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," *arXiv:1711.05101 [cs, math]*, Jan. 2019, arXiv: 1711.05101.

[15] Z. Gu, L. Gu, R. Eils, M. Schlesner, and B. Brors, "circlize implements and enhances circular visualization in R," *Bioinformatics*, vol. 30, no. 19, pp. 2811–2812, Oct. 2014.

[16] N. Gogia *et al.*, "Hippo signaling: bridging the gap between cancer and neurodegenerative disorders," *Neural Regeneration Research*, vol. 16, no. 4, pp. 643–652, Oct. 2020.

[17] S. C. D. van IJzendoorn and P. Derkinderen, "The Intestinal Barrier in Parkinson's Disease: Current State of Knowledge," *Journal of Parkinson's Disease*, vol. 9, no. Suppl 2, pp. ,323–S329, 2019.

[18] G. Chapelet, L. Leclair-Visonneau, T. Clairembault, M. Neunlist, and P. Derkinderen, "Can the gut be the missing piece in uncovering PD pathogenesis?" *Parkinsonism & Related Disorders*, vol. 59, pp. 26–31, Feb. 2019.

[19] S. Jinn *et al.*, "TMEM175 deficiency impairs lysosomal and mitochondrial function and increases alpha-synuclein aggregation," *Proceedings of the National Academy of Sciences*, vol. 114, no. 9, pp. 2389–2394, Feb. 2017, publisher: Proceedings of the National Academy of Sciences.