Explainable AI for Estimating Pathogenicity of Genetic Variants Using Large-Scale Knowledge Graphs

Shuya Abe, Shinichiro Tago, Kazuaki Yokoyama and Masaru Fuji

Daniel Gutierrez
Knight Foundation School of Computing and Information Sciences
Florida International University
Miami, USA
dguti124@fiu.edu

***Abstract*—The large amount of genomic omics data extracted from techniques like Next-Generation Sequencing necessitates the usage of AI for clinical interpretation, which can help diagnose diseases caused by genetic variants. However, the lack of explainability in traditional AI methods has led to distrust in their clinical interpretations. Original Paper Proposal: Creating an xAI system that utilizes a knowledge graph of clinical interpretations to estimate and explain the pathogenicity of a genetic variant. Current Proposal: Using a XGBoost Model trained on raw genomic/proteomic prediction scores for high estimation in distinguishing pathogenicity, with SHAP providing post-hoc explainability. Results: XGBoost and SHAP scores were compared to the performance of the paper's ”Deep Tensor” model. Conclusion: The proposed method of utilizing XGBoost improved upon the estimation performance metrics listed in the original paper but fell short of viable interpretability with SHAP. These results will help further the study of genomic medicine.**

***Index Terms*—Single Nucleotide Variant, Genomic Medicine, explainable AI, XGBoost, SHAP, Knowledge Graph**

I. Overview

Single nucleotide variants(SNV's) are mutations in the human genome that have the potential to cause autoimmune disorders and diseases like cancer. These SNV's are able to be identified and characterized by genomic omics data obtained from Next Generation Sequencing(NGS), which is a form of DNA sequencing. Due to the vast quantity and high cardinality of genomic omics data, AI is usually applied to it to accelerate the identification of genetic disorders. However, both the lack of high level explainability from traditional black box AI methods and interpretability of xAI explanations for non-technical stakeholders like physicians and healthcare administrators places a distrust in incorporating xAI within the field of Genomic Medicine.

The original paper sought to address this by proposing an xAI method that incorporated the use of knowledge graphs to power a ”Deep Tensor” classification model and text generative explainer model to achieve both high estimation and explainability performance. My proposal utilizes an XGBoost model trained on raw genomic/proteomic prediction scores to achieve high estimation, utilizing the SHAP framework to provide Post-Hoc explanations.

II. Goals

The authors of the original paper have not made their ”Deep Tensor” classification model, knowledge graph, or preprocessed datasets available to the public (this will be mentioned in-depth later in the paper). This limits the scope of this project to strictly implementing a new classification model and conducting my own exploratory data analysis, as well as data transformation on two of the three databases mentioned in the original paper. The goals are listed as followed:

- Implement an XGBoost classification model, trained on the datasets used in the assigned paper, to achieve high estimation metrics
- Compare performance metrics with those of obtained from “Deep Tensor” classification model used in the paper as well as other comparison models
- Evaluate the the interpretability and usefulness of explanations generated by SHAP and compare to that of the paper's xAI

III. Objectives

The objectives listed to achieve the goals of the proposed xAI method:

1) Obtain access to relevant datasets used in the paper(ClinVar and DbNSFP)
2) Extract, pre-process, and merge both datasets extensively.
    - Convert gzip to csv files
    - Perform Missing values and Outlier Analysis
    - Address class imbalance(more Pathogenic than Benign samples)
3) Train and Evaluate the implemented XGBoost model on the dataset.
4) Apply SHAP to the model predictions and observe local/global feature importance to achieve explainability

5) Compare the performance metrics and explainability of the proposed method to those discussed in the assigned paper.

IV. Motivation

Genomic medicine has the potential to revolutionize pathology by providing insight into the genetic background of certain diseases. Being able to accurately identify single nucleotide variants and their corresponding pathogenicity can significantly improve diagnoses, treatment, and early disease detection. Explainable AI in Genomics can increase the accuracy in estimating pathogenicity while allowing for interpretability and transparency to relevant stakeholders like non-technical clinicians and patients. This adds transparency into what are traditionally ”black box” AI methods. Hopefully the results of the assigned paper and this project will serves as a contribution to the growing usage of bleeding edge xAI in the medical field to interpret complex omics data and drive informed decision making by health professionals and any other relevant shareholders.

V. Prior Art and Challenges

A. Original Proposal

In the 2023 paper by Abe et al. ”Explainable AI for Estimating Pathogenicity of Genetic Variants using Large-Scale Knowledge Graphs”, the authors were able to implement a knowledge graph that incorporated different genomic databases to showcase the relationship between different known genetic variants and their clinical interpretations. Ontologies were used to create connections between different variant nodes and a triple-store software was utilized to visualize the entire knowledge graph(15,563,273,478 Resource Development Framework triples) as well as conduct queries on it. The knowledge graph was then used to train both the proprietary ”Deep Tensor” classification model and text generating explainer function to achieve both high estimation and explainability in pathogenic prediction. The performance of this combined xAI system was then compared to traditional methods like Decision Trees and Random Forests for metric comparison.

B. Shortcomings

There were some shortcomings that arose from the proposed XAI method in the paper. Reproducability cannot be achieved to any extent due to the ”Deep Tensor” model and knowledge graph being proprietary artifacts owned by the Japanese company Fujitsu. They also did not open-source any preprocessing techniques conducted on the large databases used to create the knowledge graphs(although the nature of knowledge graphs could indicate that prior cleaning was not necessarily required). The application of the combined XAI system in real world settings are put in doubt due to the expensive computational overhead that comes with training the classification model, performing queries on the knowledge graph, and printing textual explanations.

C. Challenges

The challenges that arose from approaching the pathogenic classification task is listed as followed:

- Learn how to preprocess both ClinVar and DbNSFP dataset without experience, considering relative “messiness” of the raw data
- Closely monitor utilization of hardware resources during extraction and pre-processing given the substantial size and dimensionality of the DbNSFP dataset
- Address Feature Selection without domain knowledge
- Interpret SHAP findings compiled from the final trained model

VI. Data Sources and Descriptions

These were the datasets used to train the proposed XAI method XGBoost. Due to the reduced scope of the project, the COSMIC database was not utilized.

A. ClinVar

- Summarized Genomic data that includes clinical phenotypes, interpretations, and descriptions for SNV’s
- Data Matrix Size: 6149023 rows x 40 columns
- Categorical feature “Clinical Significance” will be utilized in the merged dataset to serve as the class label
- Index Features kept to link ClinVar to DbNSFP: PositionVCF, ReferenceAlleleVCF, AlternativeAlleleVCF, Chromosome
- Samples missing Clinical Significance and PositionVCF values were filtered
- Cleaned data matrix size: 6149022 rows x 8 columns

B. DbNSFP

- Database that compiles functional predictions and annotations for SNV’s from various algorithms that are used to predict the variant’s inclination to cause a pathogenic behavior, such as genetic deletion and a change to protein synthesis.
- Data matrix size: 84013118 rows x 458 columns
- Only rank scores from the database’s algorithms were chosen in feature selection due to providing normalized single entry values from their corresponding raw scores.
- Index Features kept to link DbNSFP to ClinVar: Chr, pos(1-based), ref, alt
- Cleaned matrix size: 63592 rows x58 columns

VII. Methods and Tools

Itemized below are the python libraries and software used to complete this project.

A. Preprocessing

- Gzip
  - Multiple .gz files separated by chromosome number need to be combined and extracted into a single .csv file
- Dask

    - Python library that utilizes parallel processing to perform operations on data frames such as outlier analysis and feature selection. Also used to merge datasets into a single CSV file.
- Sci-kit Learn
    - Used the KNN(K-Nearest Neighbors) imputation algorithm to replace missing values and have a complete dataset.
- RowZero/CSV Editor Pro
    - GUI for viewing dataset subsets for further analysis and bug detection

B. *Model Training*

- Pandas
    - Data Manipulation for random train-test split and data frames
- Sci-Kit Learn
    - Stratified K-Fold CV
    - GridSearchCV
        - Automated tuning by training models on different hyper-parameter combinations
        - Best Model determined by K-Fold CV using training ROC AUC
- XGBoost
    - Gradient Boosted Trees
- SHAP
    - Plot summary, force, and dependency plots to highlight the global and local importance of different features in pathogenic classification

VIII. Results

A. *Best Performing Model*

The confusion matrix revealed that the model correctly classified 2456 samples as benign and 9981 samples as pathogenic in the training set, with only 90 false positives and 192 false negatives. The best average AUC across all folds was 0.99. The model exhibited high test and validation scores, low training loss, and convergence towards the end, indicating that it is not overfitting and generalizing well across all k-folds. A heat map showed a correlation between a max depth of 7 and a learning rate of 0.1, leading to a high Receiver Operating Characteristic AUC.

B. *Final Results*

Our XGBoost model achieved excellent performance on the test set, with an accuracy of 0.98, precision of 0.99, recall of 0.98, F1-score of 0.99, and ROC AUC of 0.99. The model exceeded the performance metrics of the deep tensor model (accuracy of 0.94, AUC of 0.98)as well as those of the comparison models mentioned in the assigned paper (random forests: accuracy of 0.93, AUC of 0.97; decision trees: accuracy of 0.91, AUC of 0.90). This suggests that XGBoost may be a more effective and efficient model for pathogenicity prediction.

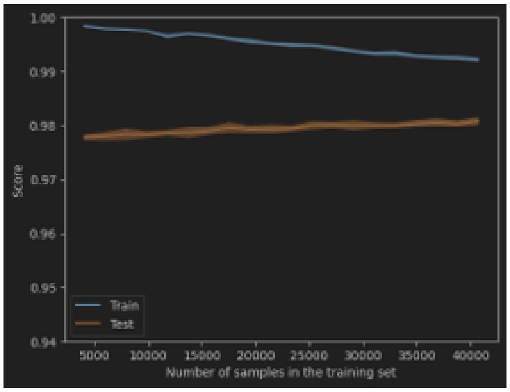

Fig. 1. Learning Curve

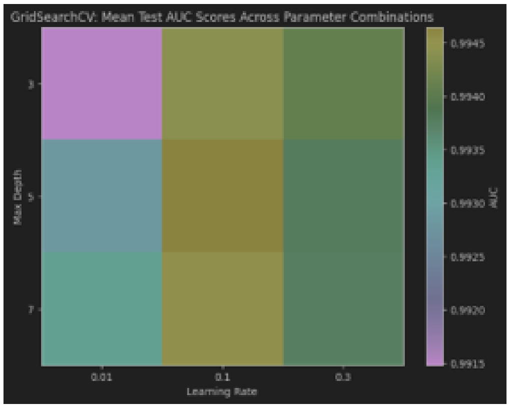

Fig. 2. Hyper-Parameter Heatmap

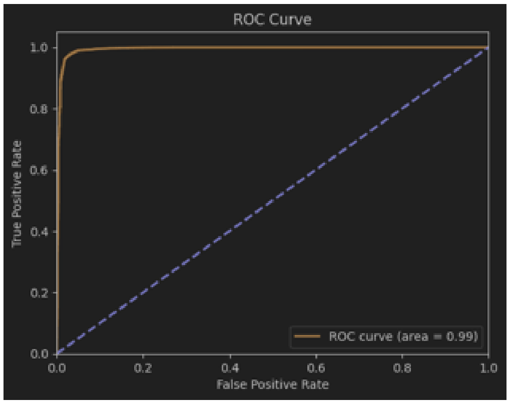

Fig. 3. Best Model ROC Curve

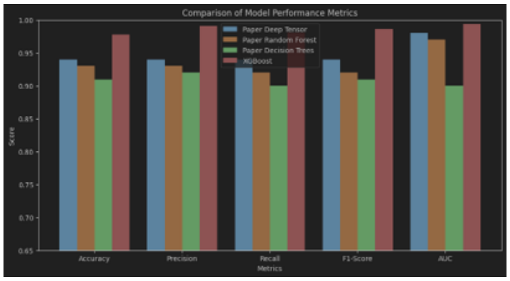

Fig. 4. Performance Metrics

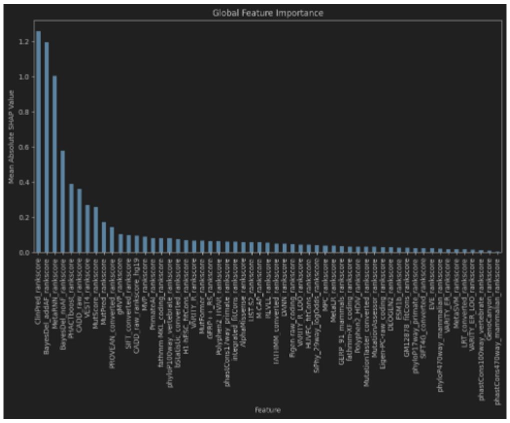

Fig. 5. SHAP Global Feature Importance

C. SHAP Values

The high accuracy of the XGBoost model indicated its potential for predicting pathogenicity of genetic variant, especially considering its improvement on the metrics of “Deep Tensor” from the assigned paper SHAP highlighted the importance of the four features ClinPred, BayesDel, and MetaRNN rankscores in classification, which equally had a hand in classifying pathogenic and benign classes. Limitations: SHAP Explanations still require domain knowledge or access to the dataset readme Potentially use OmniXAI, which offers a GPT explainer for SHAP explainations

IX. Discussion

The high accuracy of our XGBoost model (accuracy: 0.98, precision: 0.99, recall: 0.98, F1-score: 0.99, ROC AUC: 0.99) indicates its strong potential for accurately predicting the pathogenicity of genetic variants due to its significant improvement over the performance metrics of the ”Deep Tensor” model presented in the original paper. Our SHAP (SHapley Additive exPlanations) analysis revealed the importance of four key features in driving the model’s predictions: ClinPredrankscore, BayesDeladdAFrankscore, MetaRNNrankscore, and VEST4rankscore. These features exhibited both positive and negative impacts on pathogenicity classification, indicating their nuanced roles in the prediction process. However, SHAP visualizations do not provide viable interpretability to stakeholders that do not have the relevant domain knowledge needed to understand the features of the DbNSFP dataset as well as a technical understanding of the SHAP library, thus the assigned paper still performs better in explainability. This can be addressed in future work by exploring the OmniXAI library in python that offers a GPT explainer for SHAP explanations, making interpretations more human-readable.

X. Contributions and Conclusions

Explainable AI has the potential to revolutionize genomics by providing accurate and interpretable predictions. The analysis and results obtained from this experiment as well as the results from the paper represents a step towards realizing this potential and improving the diagnosis and treatment of genetic diseases

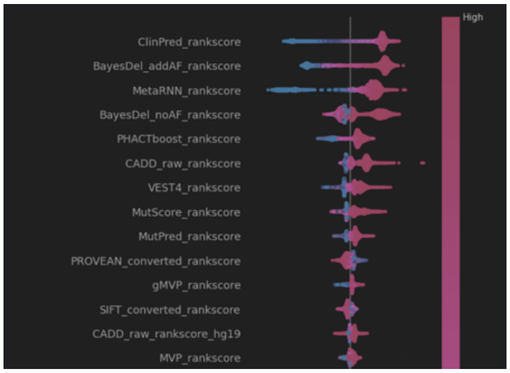

Fig. 6. SHAP Summary Plot

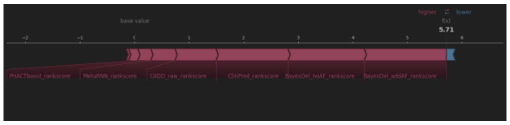

Fig. 7. SHAP Force Plot for a Sample Variant

Some contributions to Genomic Medicine that can be attributed to the results of paper includes providing a viable model for interpreting complex genomic omics data with high estimation and explainability. We can even apply this model to genetic variants listed in the DbNSFP that do not have a current pathogenic classifications assigned to it.

References

[1] Abe, S.; Tago, S.; Yokoyama, K.; Ogawa, M.; Takei, T.; Imoto, S.; Fuji, M. Explainable AI for Estimating Pathogenicity of Genetic Variants Using Large-Scale Knowledge Graphs. *Cancers* **2023**, *15*, 1118.

[2] Lundberg, S.M.; Lee, S.-I. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems* **2017**, *30*.

[3] Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* **2016**, 785–794.

[4] Liu, X.; Jian, X.; Boerwinkle, E. DBNSFP v3.0: A one-stop database of functional predictions and annotations for human non-synonymous and splice-site SNVs. *Human mutation* **2016**, *37*, 235–241.

[5] Landrum, M.J.; Lee, J.M.; Benson, M.; Brown, G.R.; Chao, C.; Chitipiralla, S.; Gu, B.; Hart, J.; Hoffman, D.; Hoover, J.; et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic acids research* **2018**, *46*, D1062–D1067.