**“SNP Based Genetic Risk Assessment Using Machine Learning”**

**By Tahsin Saira**

**Introduction**

Precision genomics-based medicine is a growing, newer idea of medicine in which a patient’s genes and DNA are analyzed and used to develop personalized treatment plans for the patient. With the help of genomic studies, modern machine learning (ML) methods have opened the door to using high dimensional inputs to predict health outcomes and accurate disease risk. Single nucleotide polymorphisms (SNPs) are the most common type of genetic variations among people, which act as biological markers, helping scientists locate genes that are the culprits behind the diseases. Polygenic risk scoring and machine learning are two primary methods used for disease risk prediction. Using a patient’s SNPs to predict individual disease risks is an essential element for precautionary medicine and disease prevention.

**Aims and Objective**

SNPs help predict an individual’s response to certain drugs, vulnerability to environmental factors such as toxins, and risk of developing diseases. SNPs can also be used to track inherited disease-associated genetic variants. Compared to polygenic risk score (PRS), machine learning algorithms have better predictive abilities for complex disease risk. Due to its dependency on linear regression, PRS have limited success in complex disease prediction.

The project aim is to analyze the prediction accuracy of an individual’s risk to a certain disease based on SNPs using PRS and various Machine learning methods. The main idea of the project is to analyze which method provides the best prediction. More research on this subject is important because measuring an individual’s susceptibility to a certain disease before invasive diagnosis may determine who will eventually come down with the disease from those who will not. Identifying the certain disease-associated SNPs that consistently distinguish disease cases from healthy controls may be particularly useful in improving risk prediction and developing individual treatment strategies.

**Literature Review**

In Genome-wide association studies (GWAS), the idea is to identify genomic variants (SNPs) using rapid scanning of markers across the complete sets of DNAs, which explains the genetic component of the observed phenotype/disease in genotyped people (Visscher et al., 2012). Once new genetic associations are identified, researchers can use the information to develop better strategies to detect, treat and prevent that disease. GWAS have already successfully identified genetic variants with significant association with complex diseases such as type 2 diabetes, Parkinson's disease, heart disorders, auto-immune diseases, prostate cancer etc. GWA studies are conducted at a population level. Genetic risk prediction models are generally done by: (1) Polygenic risk scoring (also referred to as genome-wide score); or (2) Machine learning (Abraham and Inouye, 2015).

**Polygenic Risk Scoring**

## Polygenic risk scoring uses a fixed model approach where simple statistical association testing such as linear, or logistic regression are used to estimate effect sizes from a large case, sum the contribution of a set of risk alleles to a specific complex disease (Wang et al., 2016). Polygenic risk scores can be performed weighted or unweighted. The extraction of limited numbers of SNPs from large effect sizes, over-simplifies the biological footing of any complex diseases by ignoring the larger part of the variants that may seem like non-influential but are actually making smaller individual contributions to the phenotype (Visscher et al., 2017). However, a recent study even reported that a simple Polygenic risk score outperformed machine learning methods in predicting coronary artery disease status (Gola et al 2020).

## Machine Learning Disease Prediction Models

Machine learning methods lets a set of complex statistical and computational algorithms (e.g., Support vector machine (SVM), Naive Bayes methods or Random Forest) make predictions by mathematically mapping the complex connections between a set of risk SNPs to complex disease phenotypes (Wei et al, 2009). These methods use supervised or unsupervised learnings to map the relations with complex diseases. The curse of high dimensionality problem often gives uneven performances to many state-of-art machine learning algorithms. A recent advance in machine learning is the development of deep learning algorithms that can effectively extract meaningful features from high-dimensional and complex datasets through a weighted and hierarchical learning process. Another way to evade the curse of dimensionality problem is to integrate various machine learning algorithms (Gaudillo et al 2019). A study on the integration of ML models to augment traditional methods in predicting genetic predisposition to multifactorial disease such as Asthma reported that the integrated RF-SVM model achieves the highest accuracy, precision, sensitivity. This study demonstrates the integration of ML models to augment traditional methods in predicting genetic predisposition to multifactorial diseases such as asthma (Gaudillo et al 2019). A Finnish case study implemented a gradient tree boosting method followed by an adaptive iterative SNP search to capture complex SNP-SNP interaction patterns and consequently, obtained groups of interacting SNPs, which yielded high Breast Cancer risk prediction accuracy within the SVM-based framework (Behravan et al 2018).

**Summery and Conclusion**

There are mainly two ways of conducting a risk assessment of a certain disease using a person’s DNA. One is Polygenic Risk Scoring, and another is Machine Learning. Even though in most literature, machine learning models such as Random Forest, SVM (Support vector machine) etc. outperforms PRS (depending on datasets & diseases), some studies contradict this. My project aim is to explore these methods for a certain disease of choice and conclude which method or combination of methods yields the best prediction. The SNP data will be obtained from GWAS Catalogue, openSNP database, a public repository from the 23andMe database.

A person’s genetic variants are one of the earliest and more stable indicators possible. With this tool, we can identify at-risk people early for screening, lifestyle modifications and safe non-risk people from unnecessary invasive preventions.

## References

1. Abraham, G., & Inouye, M. (2015). Genomic risk prediction of complex human disease and its clinical application. *Current Opinion In Genetics & Development*, *33*, 10-16. doi: 10.1016/j.gde.2015.06.005
2. Behravan, H., Hartikainen, J., Tengström, M., Pylkäs, K., Winqvist, R., Kosma, V., & Mannermaa, A. (2018). Machine learning identifies interacting genetic variants contributing to breast cancer risk: A case study in Finnish cases and controls. *Scientific Reports*, *8*(1). doi: 10.1038/s41598-018-31573-5
3. Gaudillo, J., Rodriguez, J., Nazareno, A., Baltazar, L., Vilela, J., & Bulalacao, R. et al. (2019). Machine learning approach to single nucleotide polymorphism-based asthma prediction. *PLOS ONE*, *14*(12), e0225574. doi: 10.1371/journal.pone.0225574
4. Gola, D., Erdmann, J., Müller‐Myhsok, B., Schunkert, H., & König, I. (2020). Polygenic risk scores outperform machine learning methods in predicting coronary artery disease status. *Genetic Epidemiology*, *44*(2), 125-138. doi: 10.1002/gepi.22279
5. Visscher, P., Wray, N., Zhang, Q., Sklar, P., McCarthy, M., Brown, M., & Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *The American Journal Of Human Genetics*, *101*(1), 5-22. doi: 10.1016/j.ajhg.2017.06.005
6. Wei, Z., Wang, W., Bradfield, J., Li, J., Cardinale, C., & Frackelton, E. et al. (2013). Large Sample Size, Wide Variant Spectrum, and Advanced Machine-Learning Technique Boost Risk Prediction for Inflammatory Bowel Disease. *The American Journal Of Human Genetics*, *92*(6), 1008-1012. doi: 10.1016/j.ajhg.2013.05.002