1. INTRODUCTION

Rheumatoid Arthritis is a complex chronic autoimmune inflammatory disease affecting 1% of the population globally (Silman & Pearson, 2002). Primarily affecting the synovial joint lining, RA begins with joint pain and stiffness as a result of inflammation. However, in its insidious nature prolonged joint inflammation may unknowingly progress to joint cartilage destruction and bone erosion, restricting movement and hence productive ability. Further it could evolve into progressive disability and a concoction of complications instigating the possibility of reducing one's life expectancy by a mean of 3-10 years (Toledano et al., 2012). Being without a cure, RA undoubtedly imposes a heavy socioeconomic burden if left undiagnosed, untreated, and undeciphered.

RA is often described to be a multifactorial disease, likely with a medley of genotypic and environmental etiologies, however, its clear pathologic mechanism remains unknown. Regardless of environmental contributions, twin studies show that the heritability of RA is estimated to be 60%, indicating a large inclination of the disease due to genetic factors (MacGregor, 2000). Current research trends in discovering susceptibility loci are dominated by genome wide association studies (GWAS) (Kim et al., 2014; Orozco et al., 2013). GWAS is an approach that examines the genetic variants across a genome of a population to identify variants associated with phenotype (Nicholls et al., 2020). However, a major hindrance for GWAS output to succeed in its clinical stage is its inability to definitively identify true causal variants that aren't riding on the associations to correlated variants due to linkage disequilibrium (Hormozdiari et al., 2015).

Most studies aimed to unravel the complexities of RA have adopted this method of genome-wide association (GWAS). Collectively these studies have reported >400 risk loci associated with RA on the EBI GWAS Catalog, however, there still remains an estimated >50% unknown genetic risk of RA (Orozco et al., 2011). Hence although having estimated that 60% of the disease liability is explained by genetic variation, >50% of theses specific disease-associated variations remain unknown. Moreover, with GWAS, the intent of search is to identify associations within the dataset and not focus on predictive performance and applications beyond that dataset. Thus, although having predictive capabilities it has never been the main purpose or strength of GWAS. In general, the fervent contributions of GWAS in identifying complex disease-associated loci have undoubtedly produced a multitude of breakthroughs in the scientific world. However, with the foundation of its method hindering the identification of true causal variants and progression to the clinical stages, its continued relevance has been put in question and it may be time to move on to improved means of discovery.

Machine learning (ML) has been a tact weapon to elucidate the data explosion that is the 21st century. With a wide array of complex algorithms, it has opened doors to its applications in various fields of science. Biological data is no exception to this technological revolution. The growing volume and complexity of biological data have encouraged the development of ML-applied methods to further the understanding of biological processes. ML is also accustomed to the incorporation of multiple variables of varying data types allowing for ease of continued enhancement of the model as more data and data types become available. To our knowledge, machine learning approaches have yet to be applied specifically to identify polymorphisms that predict RA susceptibility. However there have been machine learning approaches used to identify polymorphisms for other disease susceptibilities with results that show substantial promise (Gaudillo et al., 2013; Nebert et al., 2013). Moreover, studies with ML approaches have generally proven relevant given its clinical impacts (Nicholls et al., 2013).

As mentioned, although GWAS has been able to identify a variety of risk loci associated with RA, there still remains a large portion of RA's missing heritability. This study aims to establish a machine learning approach to identify a small subset of RA predictive polymorphisms in the hopes to elucidate RA's missing heritability. Identified SNPs are to be trained and tested for their predictive performances, measured with appropriate machine learning metrics and evaluated for its biological relevance.