TABLE I
VARIABLES DROPPED AND WHY

Variable	Description
CLASS	CLASS = 1 for all rows so it doesn't provide the machine learning model with any important information
Consequence	Redundant to the MC column
CLNDISDB	Storage in different databases is not relevant
CLNDN	ClinVar's name for information already in CLNDISDB column, redundant. Also, storage in ClinVar is not relevant to pathogenicity
CLNVI	Variant's clinical sources are not relevant to pathogenicity
CLNDISDBINCL, CLNDNINCL, CLNSIGINCL, SSR,DISTANCE, MOTIF_NAME, MOTIF_POS, HIGH_INF_POS, MOTIF_SCORE_CHANGE	Sparse Data, 0.20 percent or less of data is non-null
INTRON	Sparse Data, only 13 percent of data is non-null
CADD_RAW	Redundant, an untransformed version of CADD PHRED
BAM_EDIT	Is not relevant whether the file was edited or not
Allele	Redundant to ALT
CLNHGVS	Redundant to ALT and REF columns as well as CHROM and POS
BIOTYPE	Very low amount of variance, 48738 protein_coding and only 11 of any other type
ORIGIN	Contains values not described in the data documentation, also low variance with 47923 in one category
CLNVC	Very few values in categories other than single nucleotide variant
Feature_type	All values are uniform
CADD_PHRED, BLOSUM62,SIFT, PolyPhen	Other gene scores, not relevant to this study (deleted in a later supplementary coding file than others in this table but far before dropping nulls or encoding)