

Explainable Machine Learning for the Prioritization of Variants of Uncertain Significance in Monogenic Diabetes Genes

EBENEZER MARADESA

maradebenezer@gmail.com

University College Hospital, Ibadan <https://orcid.org/0009-0002-8458-6178>

Research Article

Keywords: monogenic diabetes, variant of uncertain significance, explainable AI, SHAP, ClinVar, variant prioritization

Posted Date: October 13th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11297260/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Explainable Machine Learning for the Prioritization of Variants of Uncertain Significance in Monogenic Diabetes Genes

Author: Ebenezer Maradesa, MBBS, Independent Researcher

Corresponding author: maradebenezer@gmail.com

October 2026

Abstract

Monogenic diabetes is caused by pathogenic variants in genes involved in pancreatic development, β -cell function, and glucose homeostasis, yet a substantial proportion of variants identified through genetic testing are classified as Variants of Uncertain Significance (VUS) due to insufficient or conflicting evidence. An explainable machine learning pipeline was developed to prioritize VUS across 22 established monogenic diabetes genes. Confidently classified pathogenic and benign single nucleotide variants (SNVs) were curated from ClinVar (GRCh38, germline, review status ≥ 1 star) and annotated with consequence, functional-prediction, conservation, and population-frequency features via the Ensembl Variant Effect Predictor (VEP) REST API and MyVariant.info (dbNSFP). Four classifiers (logistic regression, random forest, XGBoost, LightGBM) were compared via stratified cross-validation; all achieved area under the precision–recall curve (AUPRC) of 0.987–0.989, with logistic regression selected for its interpretability. On a held-out test set ($n=771$), the model achieved ROC-AUC = 0.9989 and AUPRC = 0.9970, with recall of 1.00 on the pathogenic class. Systematic feature ablation revealed that reported performance is substantially inflated by structural overlap between the feature set and American College of Medical Genetics and Genomics (ACMG) criteria used to originally classify ClinVar variants (PP3/BP4-only features: accuracy 0.953; consequence-and-frequency-only: 0.911), a circularity phenomenon previously documented for this class of model. Restricting evaluation to non-missense variants, where computational-evidence criteria could not have been directly invoked by curators, the model still substantially outperformed a majority-class baseline (99.0% vs. 87.0%), indicating genuine discriminative signal beyond simple reconstruction of curator logic. SHAP (SHapley Additive exPlanations) analysis identified CADD score, gnomAD allele frequency, and PolyPhen score as the strongest global predictors, with directionality consistent with established variant biology throughout. The trained model was applied to 6,306 VUS across these genes, producing a ranked, per-variant-explained prioritization list intended to support, not replace, expert clinical review. This work is framed explicitly as an interpretable evidence-integration and triage tool for a defined gene set, rather than as a novel independent pathogenicity discriminator, and the full pipeline, from raw ClinVar records to the final ranked output, is openly documented for reuse.

Keywords: monogenic diabetes, variant of uncertain significance, explainable AI, SHAP, ClinVar, variant prioritization

1. Background

Monogenic diabetes is caused by pathogenic variants in genes involved in pancreatic development, β -cell function, insulin secretion, and glucose homeostasis. Genetic testing plays an essential role in diagnosis, prognosis, and therapeutic decision-making for these conditions. However, a substantial proportion of variants identified through clinical sequencing are classified as Variants of Uncertain Significance (VUS) due to insufficient or conflicting evidence regarding their pathogenicity.

Existing databases such as ClinVar aggregate clinical interpretations from submitting laboratories, but a large fraction of catalogued variants remain unresolved. Numerous computational pathogenicity predictors (e.g., SIFT, PolyPhen-2, CADD, REVEL) have been developed to support variant interpretation; however, many operate as general-purpose, genome-wide tools with limited transparency regarding which factors drive any individual prediction, and limited disease- or gene-set-specific calibration.

Explainable machine learning (XML) offers an opportunity to build predictive models whose outputs can be interpreted via feature-attribution methods such as SHAP (Lundberg & Lee, 2017). Rather than replacing expert clinical interpretation, such models can serve as decision-support systems that highlight the specific evidence contributing to the prioritization of individual uncertain variants.

An important and well-documented caveat for this class of model is *circularity*: many computational predictors are themselves cited as supporting evidence within the ACMG/AMP variant classification guidelines (Richards et al., 2015), notably PP3/BP4 (computational evidence), PVS1 (loss-of-function consequence), and BA1/PM2 (population frequency). Because ClinVar labels are frequently assigned using these same criteria, models trained to predict ClinVar labels using these same categories of evidence risk partially reconstructing curator reasoning rather than discovering independent biological signal (Grimm et al., 2015). This is addressed directly through systematic ablation (Section 4.3), and the contribution of this work is framed accordingly.

2. Aim and Objectives

Aim: To develop an explainable machine learning pipeline that learns distinguishing characteristics of confidently classified pathogenic and benign variants in monogenic diabetes genes, and applies these learned patterns to prioritize variants of uncertain significance for further investigation.

Objectives: 1. Curate ClinVar records for 22 genes associated with monogenic diabetes. 2. Construct a variant-level, quality-filtered dataset (GRCh38, germline, single nucleotide variants, review status ≥ 1 star). 3. Annotate variants with consequence, population frequency, conservation, and functional-prediction features. 4. Train and compare supervised classifiers on confidently classified (pathogenic/benign) variants. 5. Characterize the extent of label circularity via systematic feature ablation. 6. Interpret model predictions using SHAP, at both global and per-variant level. 7. Apply the trained model to VUS and produce a ranked, explained prioritization list. 8. Provide a reproducible, publicly documented pipeline.

3. Methods

3.1 Data source and gene selection

Variant records were obtained from ClinVar (variant_summary.txt), downloaded 30 July 2026. Records were filtered to 22 genes with established roles in monogenic diabetes: *ABCC8*, *LMNA*, *WFS1*, *GCK*, *HNF1A*, *EIF2AK3*, *HNF1B*, *INSR*, *HNF4A*, *GATA6*, *FOXP3*, *RFX6*, *BLK*, *SLC2A2*, *CEL*, *KCNJ11*, *AKT2*, *KLF11*, *NEUROD1*, *PPARG*, *PDX1*, *PAX4*, *INS*. This yielded 38,204 raw ClinVar records.

3.2 Variant filtering and cohort definition

Records were filtered sequentially:

1. **Genome build:** restricted to GRCh38 (removing duplicate GRCh37 records of the same variant; 86 GRCh37-only variants with no GRCh38 counterpart were excluded rather than lifted over, given their negligible proportion of the dataset, ~0.5%).
2. **Origin:** restricted to germline variants (OriginSimple == 'germline'), excluding somatic, unknown, and not-applicable origin records.
3. **Clinical significance:** variants were assigned to one of two cohorts:
 - **Training cohort:** Pathogenic, Likely pathogenic, Pathogenic/Likely pathogenic (pooled as “pathogenic-like”, label = 1) versus Benign, Likely benign, Benign/Likely benign (pooled as “benign-like”, label = 0). Variants with conflicting classifications, risk-allele designations, or non-diagnostic annotations (e.g., “association”, “drug response”) were excluded from training.
 - **VUS cohort:** variants labeled exactly Uncertain significance, retained for downstream prioritization but excluded from model training.
4. **Review status:** both cohorts were restricted to ClinVar review status of “criteria provided, single submitter” or higher (≥ 1 star), excluding variants with no assertion criteria or conflicting-classification review status.
5. **Variant type:** restricted to single nucleotide variants (SNVs), as conservation and functional-prediction scores used as model features are not well-defined for structural variants (deletions, duplications, indels). This is stated as an explicit scope limitation (Section 6).
6. **Biotype:** one variant whose canonical transcript was annotated as a long non-coding RNA (lncRNA) rather than protein-coding was excluded as an atypical edge case inconsistent with the coding-variant-focused annotation strategy.

Prior to the SNV/biotype filters, GRCh38 germline records numbered 8,995 (training) and 6,660 (VUS). Restricting to SNVs and excluding one lncRNA-biotype and seven non-canonical-transcript outliers produced a final training cohort of **7,706 labeled SNVs** (benign-like: n=5,690, 73.8%; pathogenic-like: n=2,016, 26.2%) and a VUS cohort of **6,306 SNVs**.

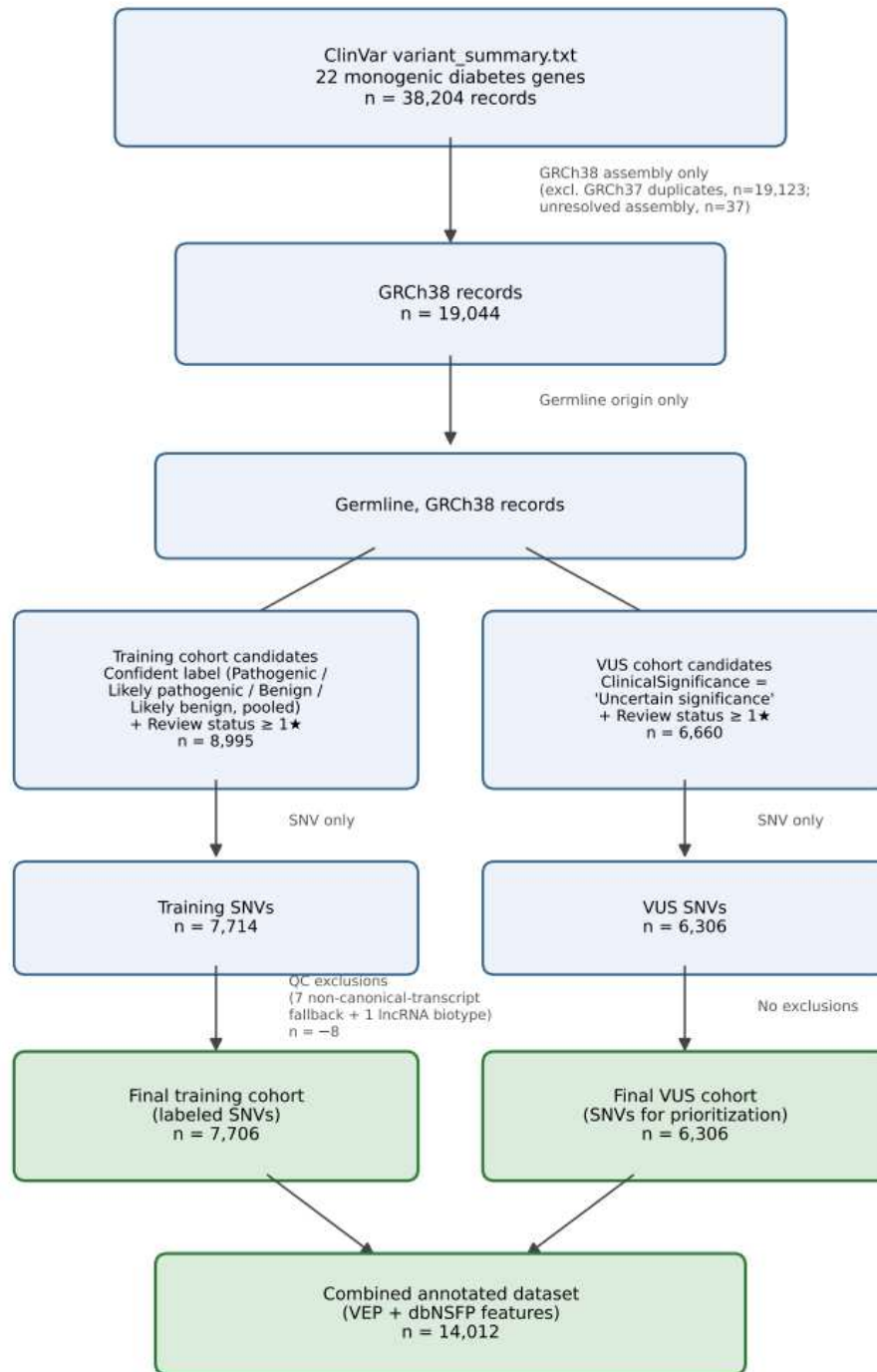


Figure 1: cohort filtering flowchart, raw ClinVar diabetes-gene records (n=38,204) → GRCh38 only (n=19,044) → germline only → confident-label training / confident-review VUS split (n=8,995 / n=6,660) → SNV-only (n=7,714 / n=6,306) → final QC exclusions → final training (n=7,706) / VUS (n=6,306) cohorts.

3.3 Variant annotation

Variants were annotated using two complementary, publicly available REST services rather than a local Ensembl VEP installation, prioritizing accessibility and connectivity resilience over the marginal reproducibility advantage of a version-pinned local cache; the specific release/version queried is recorded for reproducibility (Section 7).

- **Ensembl VEP REST API** (release **116**, confirmed via `data/vep_rest_raw_responses/release_info.json`; queried 12th August 2026), batched in groups of 200 variants via the `/vep/homo_sapiens/region` POST endpoint, with `canonical=1` to identify one representative transcript per variant. Extracted fields: most severe consequence, impact category, SIFT score/prediction, PolyPhen score/prediction, and gnomAD exome allele frequency (via colocated variant records).
- **MyVariant.info** (dbNSFP-backed), queried [CONFIRM date], batched in groups of 1,000 variants via the `/v1/variant` POST endpoint. Extracted fields: GERP++ rejected-substitutions score (`gerp++.rs`), phyloP 100-way vertebrate conservation rankscore, CADD Phred score, and REVEL score.

For both services, only an explicit whitelist of biologically motivated fields was extracted; ClinVar-derived annotations returned alongside these responses (e.g., `clin_sig`, ClinVar/OMIM cross-references, `phenotype_or_disease` flags) were deliberately excluded from the feature set to prevent label leakage, since these fields directly or indirectly echo the classification being predicted.

All raw API responses were persisted to disk in batch-level JSON files prior to parsing, to support reproducibility and auditability independent of live API availability.

3.4 Feature engineering

Twelve features were used for modeling, spanning four evidentiary categories:

Category	Features	ACMG criterion overlap
Consequence/impact	most_severe_consequence (one-hot encoded), impact_order (ordinal: MODIFIER=0 ... HIGH=3)	PVS1
Population frequency	gnomad_af_filled (missing values imputed to 0, log10-transformed, standardized)	BA1/BS1, PM2
Computational prediction	sift_score_filled, polyphen_score_filled, cadd_phred_filled, revel_score_filled (sentinel-imputed at -1 where structurally undefined)	PP3/BP4
Conservation	gerp_rs_filled, phylop_rankscore_filled (sentinel-imputed at -1)	PP3/BP4

Each imputed continuous feature was paired with a binary missingness indicator (has_sift_score, has_polyphen_score, has_revel_score, has_dbnsfp_coverage), since missingness in these predictors is structurally informative (e.g., SIFT/PolyPhen/REVEL are defined only for missense variants; dbNSFP conservation/CADD annotations were empirically found to be populated predominantly for nonsynonymous coding variants, present for 87.9% of missense variants versus a small minority of synonymous/intronic/UTR variants).

Gene identity (GeneSymbol) was deliberately excluded as a model feature, despite its availability, to avoid the model learning gene-specific label-frequency shortcuts (e.g., one gene in our cohort, *ABCC8*, was disproportionately represented in the training set relative to the VUS set: 18.3% vs. 9.0%) rather than transferable biological signal.

All learned transforms (one-hot encoding, standard scaling, log transformation) were implemented within a scikit-learn ColumnTransformer/Pipeline, fit exclusively on the training split, to prevent cross-fold and train/test leakage. Deterministic transforms (missingness-sentinel imputation, ordinal mapping) were applied prior to pipeline construction.

3.5 Model training and comparison

The labeled cohort was split 90/10 (train/test), stratified by label. Four classifiers were compared via 5-fold stratified cross-validation on the training split: logistic regression, random forest, XGBoost, and LightGBM, each with class-imbalance correction (class_weight='balanced' or equivalent scale_pos_weight). Primary model-selection metric was AUPRC, given the ~2:1 class imbalance.

3.6 Interpretation

SHAP values were computed using `shap.LinearExplainer` (exact, not approximate, for the selected linear model), with the training set as the background distribution and the held-out test set used for global summary interpretation.

3.7 VUS prioritization

The final fitted pipeline was applied to the 6,306-variant VUS cohort to generate a predicted pathogenicity probability per variant. SHAP values were computed per VUS, and the top three contributing features (by absolute SHAP magnitude) were recorded alongside each prediction to provide a per-variant rationale.

4. Results

4.1 Model comparison

Table 1. Five-fold stratified cross-validation results (training split only, n=6,935).

Model	ROC-AUC		AUPRC		F1	Precision	Recall	Accuracy
Logistic Regression	0.9939	± 0.0033	0.9890	± 0.0043	0.9701	0.9657	0.9746	0.9843
XGBoost	0.9928	± 0.0035	0.9876	± 0.0049	0.9681	0.9655	0.9708	0.9833
LightGBM	0.9932	± 0.0034	0.9873	± 0.0046	0.9670	0.9639	0.9702	0.9827
Random Forest	0.9922	± 0.0040	0.9868	± 0.0062	0.9663	0.9598	0.9730	0.9823

All four classifiers performed within a narrow band (AUPRC 0.9868–0.9890; ROC-AUC 0.992–0.994), with differences smaller than the cross-fold standard deviation, indicating no meaningful advantage for more flexible (tree-based) models over logistic regression on this feature set, consistent with a largely additive relationship between features and label. Logistic regression was selected as the primary model for its exact, fast SHAP interpretability (`LinearExplainer`) and marginal leading performance.

4.2 Held-out test set performance

On the held-out test set (n=771):

Class	Precision	Recall	F1
Benign-like (0)	1.00	0.97	0.99
Pathogenic-like (1)	0.93	1.00	0.96

ROC-AUC = 0.9989; AUPRC = 0.9970. Recall on the pathogenic-like class was 1.00, at a modest precision cost (0.93), a trade-off we consider appropriate for a triage tool, where failing to flag a truly pathogenic variant is a more consequential error than an unnecessary flag.

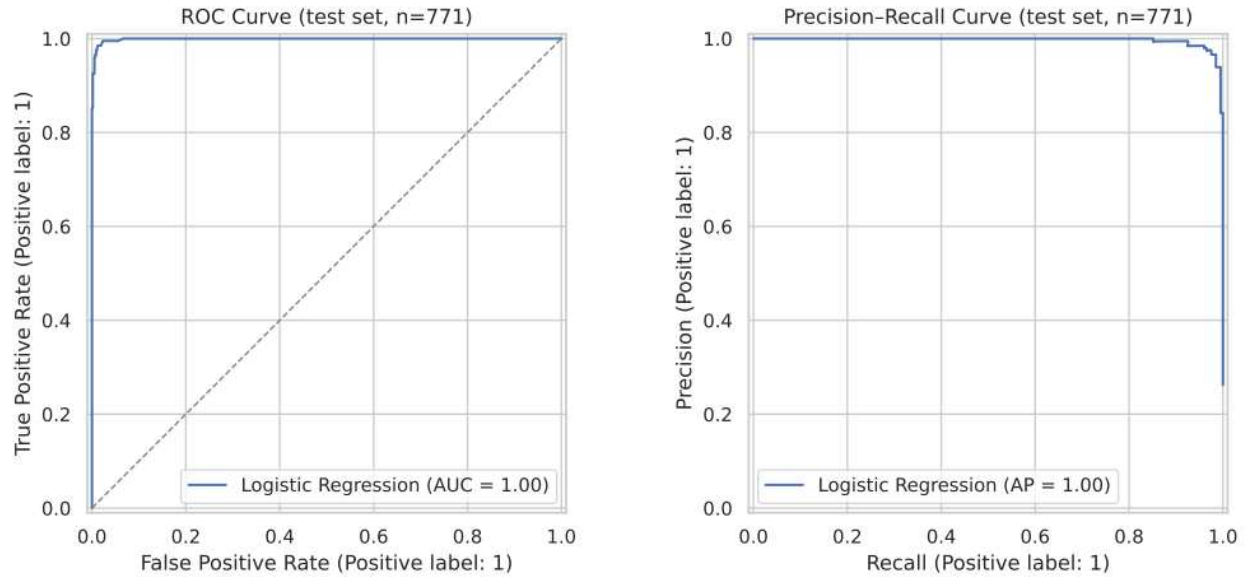


Figure 2: ROC curve and precision–recall curve for the held-out test set, side by side.

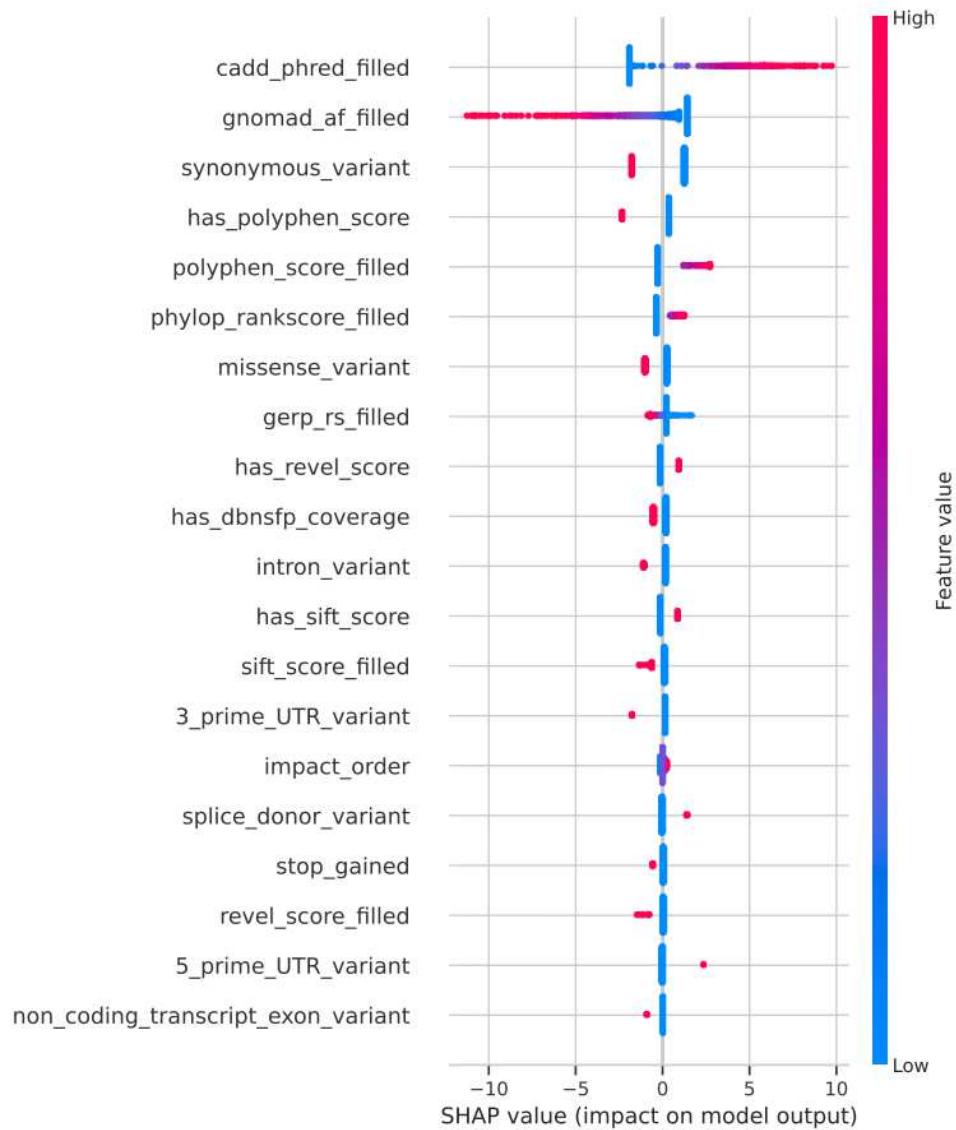
4.3 Feature ablation and circularity characterization

To characterize the extent to which model performance reflects genuine independent biological signal versus overlap with ACMG classification criteria, we conducted systematic ablations:

Feature subset	Test accuracy
Full feature set	0.9896
PP3/BP4-only (SIFT, PolyPhen, CADD, REVEL, conservation)	0.9533
PVS1/BA1-only (consequence, impact, allele frequency)	0.9105
Full model, evaluated on non-missense subset only (n=614)	0.9902 (vs. 0.8697 majority-class baseline)

No individual feature category was solely responsible for the high full-model performance; each category (consequence/impact, population frequency, computational prediction, conservation) was independently a strong predictor of ClinVar label, consistent with each corresponding to a distinct ACMG criterion category also used by curators. Restricting evaluation to non-missense variants, where PP3/BP4 computational evidence could not have been directly invoked by curators, the model still substantially outperformed a majority-class baseline (99.0% vs. 86.97%), suggesting the model captures genuine discriminative signal beyond simple reconstruction of PP3/BP4 evidence, even as PVS1/BA1-adjacent circularity likely remains present in this subgroup.

4.4 Global feature importance (SHAP)



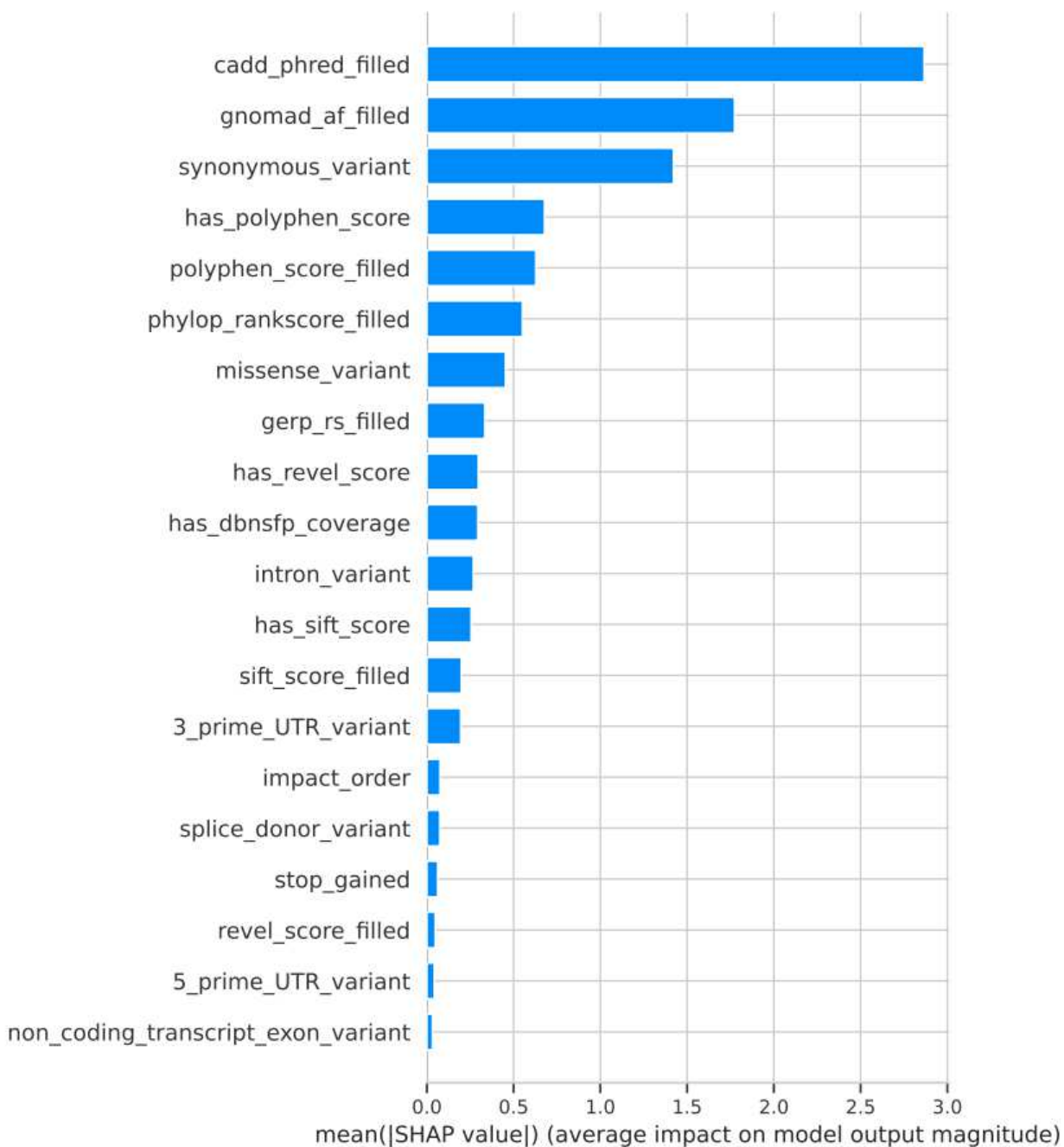


Figure 3: SHAP summary plot, two panels: (a) beeswarm/dot summary plot showing per-sample SHAP value distribution and feature-value coloring, (b) bar plot of mean absolute SHAP value per feature. Both generated via `shap.summary_plot(shap_values.values, X_test_transformed, feature_names=clean_names)` and `plot_type='bar'` respectively, computed on the held-out test set ($n=771$).

Ranked by mean absolute SHAP value: CADD Phred score (2.87), gnomAD allele frequency (1.77), synonymous-consequence indicator (1.42), PolyPhen score (0.63), and phyloP conservation rankscore (0.55) were the strongest global contributors. Directionality was biologically consistent throughout: higher allele frequency reduced predicted pathogenicity;

higher CADD/PolyPhen scores and lower (more deleterious) SIFT scores increased it; synonymous consequence strongly reduced predicted pathogenicity, consistent with its near-exclusive (99.88%) benign-like labeling in the training data.

4.5 VUS prioritization output

Applying the trained pipeline to the 6,306-variant VUS cohort yielded a predicted pathogenicity probability distribution with mean 0.694 (SD 0.349), median 0.873, interquartile range [0.439, 0.980] (Figure 4). 47.08% and 72.66% of VUS were classified at a pathogenicity probability greater than 0.9 and 0.5 respectively. The highest-ranked variants spanned multiple genes (*GCK*, *HNF1A*, *LMNA*, *WFS1*, *PDX1*, *ABCC8*, *INS*, *BLK*, *RFX6*, *HNF4A*), with CADD score, PolyPhen score, and gnomAD allele frequency recurring as the dominant SHAP drivers among top-ranked candidates (Table 2).

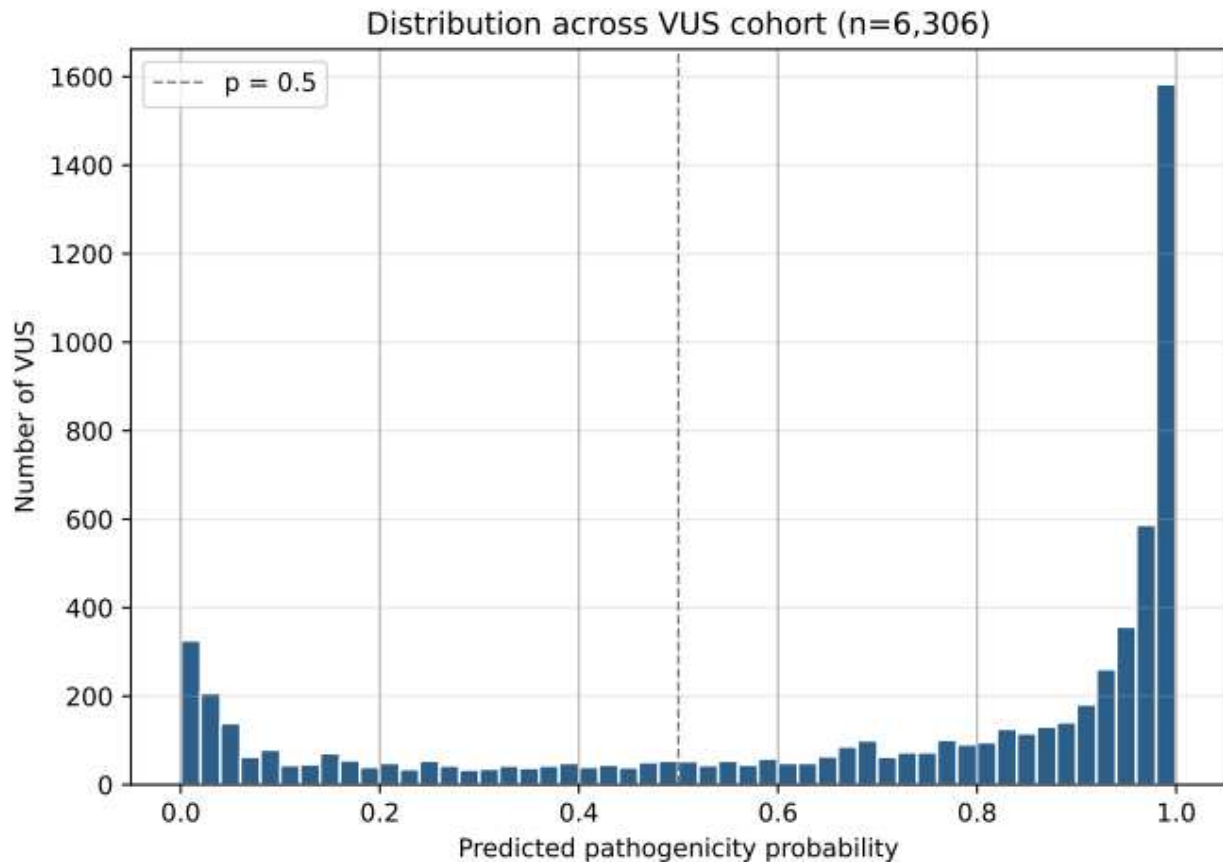


Figure 4: histogram of `predicted_pathogenicity_prob` across all 6,306 VUS, generated via `vus_ranked['predicted_pathogenicity_prob'].hist(bins=50)`.

Table 2. Top 10 prioritized VUS by predicted pathogenicity probability.

Rank	VariationID	Gene	P(pathogenic)	Top SHAP drivers
1	3233734	GCK	0.9999	CADD (6.02), PolyPhen
2	4743118	GCK	0.9999	CADD (5.80), PolyPhen
3	519299	LMNA	0.9998	CADD (5.80), PolyPhen
4	1744462	HNF1A	0.9998	CADD (5.35), PolyPhen
5	4728583	INS	0.9998	CADD (8.74), gnomAD AF
6	4277851	WFS1	0.9997	CADD (8.29), gnomAD AF
7	4763860	PDX1	0.9997	CADD (8.51), synonymous
8	556210	ABCC8	0.9997	CADD (6.25), gnomAD AF
9	36828	HNF1A	0.9996	CADD (6.25), PolyPhen
10	4741362	HNF1A	0.9996	CADD (6.48), PolyPhen

Full table available as supplementary material

5. Discussion

This work demonstrates that an interpretable logistic-regression-based pipeline, using conventional, publicly accessible variant-annotation resources, can integrate multiple lines of computational evidence into a single, ranked, and per-variant-explained prioritization output for VUS in monogenic diabetes genes. Four classifiers spanning linear and tree-based architectures converged to a narrow performance band, suggesting the discriminative signal available from these features is largely additive rather than dependent on complex feature interactions, a finding that itself supports the choice of a simple, fully interpretable model over a more flexible one, since no accuracy is sacrificed for the interpretability gained.

Systematic ablation places an important caveat on the near-ceiling reported accuracy. Because ClinVar labels are frequently assigned using ACMG criteria that draw on the same evidence categories used as model features, PVS1 for loss-of-function consequence, PP3/BP4 for computational predictors, BA1/PM2 for population frequency, a substantial portion of apparent model performance reflects reconstruction of existing curator logic rather than novel biological discovery. This circularity is a structural property of using ClinVar as ground truth for any variant-effect classifier trained on standard annotation features, not a defect specific to this pipeline, and has been previously documented for missense-effect predictors more broadly (Grimm et al., 2015). Critically, however, restricting evaluation to the subset of variants for which the dominant computational-evidence features (SIFT, PolyPhen, REVEL) are structurally undefined, non-missense variants, where PP3/BP4 could not have been directly invoked by a curator, the model still achieved 99.0% accuracy against an 87.0% majority-class baseline. This indicates that discriminative signal beyond reconstruction of PP3/BP4-style evidence is genuinely present, even as circularity linked to PVS1 (consequence severity) and BA1/PM2 (allele frequency) likely persists within this subgroup, since variant_summary.txt does not record which specific criteria a given submitter invoked.

The primary contribution of this work is therefore framed not as a novel independent pathogenicity discriminator, but as: (1) a disease-gene-set-specific integrator and triage tool, applying existing computational evidence consistently and at scale to variants that have not yet received expert curation; and (2) a demonstration of per-variant explainability, via SHAP, as a mechanism for making that triage output auditable and clinically interpretable rather than a black-box probability alone. This distinction matters practically: a triage tool that reliably reconstructs the logic a curator would apply, and can show its reasoning per variant, has direct utility in reducing the manual burden of VUS review, independent of whether it also discovers biology no curator has already considered.

Several design choices reflect this framing. Gene identity was deliberately withheld as a feature to prevent the model from learning gene-specific label-frequency shortcuts, at some cost to potential sensitivity for genes with strong dosage-sensitivity effects not otherwise captured by the included annotation features (Section 6); this trade-off favors a model whose reasoning generalizes across the gene set over one that is more accurate but less transparent about why. Similarly, the asymmetric precision–recall trade-off observed on the test set (recall = 1.00, precision = 0.93 on the pathogenic-like class) is treated as a favorable operating point rather than a shortcoming, since under-flagging a true pathogenic variant is a more consequential error than an unnecessary flag in a triage context feeding into expert review, not autonomous clinical decision-making.

The prioritized VUS output itself illustrates the intended use case: high-ranked candidates spanned multiple genes rather than concentrating in one, and were consistently associated with the same evidence categories (CADD, PolyPhen, gnomAD allele frequency) identified as globally important, giving a reviewer a consistent, auditable basis for triage rather than an opaque score. Future clinical or research use of this ranked output should treat it as a prioritization aid for directing limited curation effort, not as a substitute for functional validation or multidisciplinary variant-classification review.

6. Limitations

- ClinVar classifications evolve over time and may change as new evidence becomes available; the training labels represent a single snapshot.
- Structural circularity exists between the feature set and ACMG classification criteria (Section 4.3); reported performance should not be interpreted as a measure of independent discriminative novelty.
- `variant_summary.txt` does not record which specific ACMG criteria were invoked by individual submitters, precluding a fully precise separation of circular from independent evidence.
- Analysis was restricted to single nucleotide variants; structural variants (deletions, duplications, indels) were excluded, as key features (conservation and functional-prediction scores) are not well-defined for these variant classes.
- CADD/REVEL/GERP/phyloP annotation was obtained via a REST API (MyVariant.info/dbNSFP) rather than a version-pinned local installation; while raw responses were archived for reproducibility, live re-querying may return different results as underlying databases update.

- Functional validation of prioritized VUS is beyond the scope of this project; predicted pathogenicity probabilities are intended to support, not replace, expert clinical and functional assessment.
- Gene identity was excluded as a feature to avoid gene-specific label-frequency shortcuts; this may reduce sensitivity to true gene-specific constraint (e.g., dosage sensitivity) not otherwise captured by the included features.

7. Reproducibility and Data/Code Availability

Code, notebooks, and pipeline documentation are available at: <https://github.com/maradeben/clinvar-vus-diabetes>

- ClinVar source: variant_summary.txt, downloaded 30 July 2026
- Ensembl VEP REST API: **release 116** (see data/vep_rest_raw_responses/release_info.json), queried 12 August 2026; raw batch responses archived at data/vep_rest_raw_responses/
- MyVariant.info (dbNSFP-backed): queried 30 August 2026; raw batch responses archived at data/myvariant_raw_responses/
- Python environment: Python 3.13.9
- Trained model artifact: models/logreg_pipeline_v1.joblib, with accompanying metadata (models/logreg_pipeline_v1_metadata.json)
- Final VUS prioritization output: outputs/vus_prioritized_ranking.csv

Because the Ensembl VEP REST API reflects Ensembl's current release at query time rather than a version a reader can permanently pin, the archived raw JSON batch responses above, not a live re-query, constitute the reproducible record of the annotations used in this analysis.

Declarations

Ethics approval and consent to participate

This study used only publicly available, de-identified, aggregate variant data from ClinVar and did not involve human subjects, patient contact, or the collection of new clinical data. No institutional ethics or IRB approval was required.

Data availability

All data used are derived from ClinVar (publicly available; <https://www.ncbi.nlm.nih.gov/clinvar/>), Ensembl VEP, and MyVariant.info (dbNSFP), all of which are freely accessible. Processed datasets, code, and the final prioritized VUS output are available at the GitHub repository above.

Competing interests

The author declares no competing interests.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author contributions

E.M. conceived the study, designed and implemented the analysis pipeline, performed the analysis, and wrote the manuscript.

Use of AI tools. Large language model assistance (Claude, Anthropic) was used during pipeline development and manuscript drafting, including code review, debugging support, and editorial suggestions. All analytical decisions, code execution, and interpretation of results were performed and verified by the author.

Consent for publication

Not applicable.

8. Future Directions

- Temporal validation using historical ClinVar releases, assessing whether high-ranked VUS in an earlier release were subsequently reclassified as pathogenic.
- Incorporation of protein structural/domain-level features and gene-level constraint metrics (e.g., gnomAD pLI/LOEUF) as an alternative to raw gene identity.
- Extension to structural variants via a separate, appropriately scoped feature set.
- Benchmarking against established meta-predictors (e.g., REVEL, BayesDel) directly on this gene set.
- Development of an interactive interface for clinical exploration of ranked VUS and their SHAP explanations.

References

1. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*. 2015;17(5):405-424.
2. Grimm DG, Azencott CA, Aicheler F, et al. The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Hum Mutat*. 2015;36(5):513-523.
3. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2017;30:4765-4774.
4. Landrum MJ, Lee JM, Benson M, et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res*. 2018;46(D1):D1062-D1067.
5. McLaren W, Gil L, Hunt SE, et al. The Ensembl Variant Effect Predictor. *Genome Biol*. 2016;17(1):122.
6. Liu X, Jian X, Boerwinkle E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum Mutat*. 2011;32(8):894-899.
7. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 2020;581(7809):434-443.

8. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016:785-794.
9. Ke G, Meng Q, Finley T, et al. LightGBM: A highly efficient gradient boosting decision tree. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2017;30:3146-3154.
10. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. *J Mach Learn Res*. 2011;12:2825-2830.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [vusprioritizedranking.csv](#)