An Actor-Critic Reinforcement Learning Framework for Variant Evidence Interpretation

Dylan Lawless*[1]

[1]Department of Intensive Care and Neonatology, University Children's Hospital Zürich, University of Zürich, Switzerland.

March 14, 2025

Acronyms

ACE	associative search element	3
ASE	adaptive critic element	3
AUC	area under the curve	2
ROC	receiver operating characteristic	2
QV	Qualifying variant	2
RL	reinforcement learning	2
TD	temporal-difference	5

Project code: https://github.com/DylanLawless/rl2025lawless

*Addresses for correspondence: Dylan.Lawless@uzh.ch

Abstract

We present a reinforcement learning (RL) framework that uses established genomic metrics – such as the GuRu score, variant/gene risk priors, and population frequency – to estimate the probability of observing a given genetic variant in disease. Importantly, our approach does not directly predict variant pathogenicity; instead, it quantifies the cumulative evidence supporting a variant's clinical observability within a Bayesian context. Using simulated genetic data with a range of variability and label noise, we systematically evaluated the actor-critic algorithm performance across multiple scenarios, employing metrics including receiver operating characteristic (ROC) curves, area under the curve (AUC) , calibration plots, and learning dynamics. Results indicate predictive accuracy and effective learning, demonstrating RL's potential as a practical tool for genomic variant interpretation, setting the stage for integration into a broader Bayesian classification framework.

1 Introduction

Precise interpretation of genetic variants remains a central challenge in precision medicine, significantly impacting clinical decision-making and patient care. Differentiating pathogenic from benign variants accurately is essential for genetic diagnostics. Standard classification methods often face limitations due to incomplete or uncertain annotations, motivating the exploration of adaptive machine learning techniques. To address these challenges, we simulate genomic data that captures the main target features of evidence underlying variant interpretation and the inherent uncertainties encountered in clinical settings.

Qualifying variant (QV)s represent genomic alterations identified through stringent criteria in genomic processing pipelines and form the foundation for calculating the GuRu score. In our simulation, each QV is assigned a GuRu score based on the cumulative tally of effect evidence, comprising functional assays, computational predictions, and clinical observations, thereby reflecting the strength of evidence for a variant's potential pathogenicity. The selection and classification of these QVs adhere to established best practices in variant reporting and analysis (1–5), as well as standardised workflows (6–8). Furthermore, population allele frequencies from databases such as gnomAD (9) inform the estimation of variant probabilities, while resources like ClinVar (10) and ClinGen (11) provide clinical evidence that reinforces the interpretation of variant impact and associated phenotypes. Since this dataset is inherently noisey, we begin by substituting with simpler synthetic data, laying the groundwork for future application of these methods to empirical genomic datasets.

RL provides an appealing alternative to traditional supervised methods by utilising evaluative rather than instructive feedback. Instead of explicitly labelled outcomes, RL algorithms receive scalar rewards reflecting decision quality, thereby naturally balancing exploration of uncertain genomic space and exploitation of known information. Inspired by the classical k-armed bandit problem, our approach adapts an actor-critic algorithm to genomic data, using key features already available clinically – such as the GuRu score, gene risk categorisation, and population allele frequency – to classify the probability of a variant being assigned as pathogenic based on the existing evidence. This is subtly but importantly distinct form assigning pathogenicity itself, which is a separate task.

We recall the actor-critic architecture of the pole-balancing control problem as described by Barto et al. (12, 13). In that work, a two-component adaptive system was introduced, comprising an adaptive critic element (ASE) (associative search element) and an associative search element (ACE) (adaptive critic element), designed to solve the challenging pole-balancing control problem without any prior knowledge of the system dynamics. The system learns exclusively from sparse, delayed failure signals. Specifically, the ASE employs a stochastic, reinforcement-based update rule given by

$$w_i(t+1) = w_i(t) + \alpha \, r(t) \, e_i(t),$$

where $w_i(t)$ denotes the weight of the ith input connection at time t, α is the actor's learning rate, $r(t)$ is the reinforcement signal provided by the environment at time t, and $e_i(t)$ is the eligibility trace capturing the recent history of activity for that connection. The eligibility trace itself decays over time according to

$$e_i(t+1) = \delta \, e_i(t) + (1-\delta) \, y(t) \, x_i(t),$$

with δ representing the decay factor, $y(t)$ the output of the system at time t, and $x_i(t)$ the ith component of the input vector. Meanwhile, the critic's weight for the ith input is denoted by $v_i(t)$ and is updated via

$$v_i(t+1) = v_i(t) + \beta \, [r(t) + \gamma \, p(t) - p(t-1)] \, x_i(t),$$

where β is the critic's learning rate, γ is the discount factor determining the influence of future rewards, and $p(t)$ is the critic's current prediction of the cumulative future reward. This update rule allows the critic to refine its predictions of long-term rewards by comparing the current prediction with the observed reinforcement.

The architecture addresses the credit-assignment problem by enabling the system to learn which actions contribute to success or failure in a complex, uncertain environment. In our study, we extend this actor-critic framework to a genomic setting by integrating established genomic metrics (e.g., the GuRu score, variant/gene risk priors, and population frequency). Instead of directly predicting variant pathogenicity, our RL method estimates the probability of observing disease-associated variants by updating weights via the temporal-difference (TD) error,

$$\delta(t) = r(t) + \gamma\, p(t) - p(t-1).$$

This approach quantifies the cumulative evidence that supports a variant's clinical observability within a Bayesian context.

We systematically quantify the performance of our actor–critic RL method using simulated genomic scenarios that incorporate data imperfections such as label noise. Our goal is to identify RL methodologies and optimal parameter configurations that deliver robust, accurate predictive capabilities. These findings will form the foundation for integrating our RL framework into a broader Bayesian analytical approach for interpreting genetic evidence in clinical diagnostics.

2 Methods

Our investigation employed a synthetic dataset designed to mimic the characteristics of genetic variant data. The dataset comprised 2,000 variants, with 50% of the entries containing known pathogenicity labels and the remaining 50% left unlabelled for prediction purposes. Variants were generated to reflect realistic genomic scenarios by incorporating features such as the GuRu score, gene number, and population frequency. The variant distributions for each feature were designed using a mixture of priors. Variants in genes numbered 4 to 10 were assigned higher prior probabilities of pathogenicity, while variants in genes numbered 1 to 6 had lower pathogenicity priors. Genes 4–6 appeared in both categories, reflecting realistic cases where a single gene can harbour both pathogenic and benign variants depending on other genomic features. Furthermore, to simulate realistic errors and uncertainties commonly encountered in empirical genomic datasets, we intentionally introduced misannotations by randomly flipping the true pathogenicity labels for a predefined proportion (10%, 20%, and 30%) of the variants.

The GuRu score, also previously reported as the ACMGuru score, is a composite metric designed to quantify the totality of the best-known evidence regarding a genetic variant's function, particularly its potential to be classified as pathogenic or benign. Rather than providing a direct measure of pathogenicity, the GuRu score aggregates diverse types of evidence – including clinical data, functional assays, and computational predictions – to reflect the amount of prior knowledge available about a variant. This evidence-based measure helps to objectively gauge the consensus on a variant's classification and can be substituted by other similar metrics that integrate multiple lines of evidence, thereby offering a flexible tool for variant interpretation in genomic studies.

The reinforcement learning framework was implemented via an actor-critic algorithm. In our formulation, the state space was constructed by discretising the GuRu score into four bins, the population frequency into three bins, and the gene risk into a binary indicator; the product of these discretisations yielded 24 unique states. The action space was binary, corresponding to the two possible classifications: benign (0) or pathogenic (1). At each iteration, the RL agent observed a state corresponding to a variant and selected an action using a probabilistic policy derived from a sigmoid function applied to a vector of actor weights. A reward of +1 was granted if the predicted label matched the true pathogenicity, otherwise a penalty of -1 was imposed. The temporal-difference (TD) error, defined as the difference between the received reward and the critic's estimate of the state value, was used to update both the actor and critic weights using learning rates α and β, respectively.

In order to assess the robustness of our model, we varied three critical parameters: the noise level in the training labels (η), the actor learning rate (α), and the critic learning rate (β). For each combination of these parameters, the model was trained for 20 epochs, and a variety of performance metrics were recorded.

The evaluation comprised epoch-level measures of average TD error and average reward, as well as more detailed assessments including ROC curves with associated AUC values, precision, recall, F1 scores, cumulative learning curves, and calibration plots. All these metrics were aggregated and visualised for comparison of the effects of different parameter settings.

To improve computational efficiency and allow extensive hyperparameter evaluation, we employed parallel computing through the `doParallel` and `foreach` packages in R. We used multiple cores to simultaneously evaluate different parameter combinations, reducing the computational time required for extensive simulations. Worker-specific logging, identified by process IDs, ensured clear tracking and reproducibility of parallel computations.

Synthetic data were generated according to the specified stratification and noise levels, and the RL model was trained on a randomly partitioned training set for each combination of parameters. Performance on a held-out test set was evaluated and visualised. Additionally, we generated visualisations of feature distributions, correlation matrices, and covariance matrices across noise levels, providing comprehensive diagnostic insights into the synthetic dataset's structure and variability.

In our actor-critic RL system, the key variables are defined as follows. The actor's weight for the ith connection at time t is denoted by $w_i(t)$ and is updated according to

$$w_i(t+1) = w_i(t) + \alpha\, r(t)\, e_i(t),$$

where α is the actor's learning rate, $r(t)$ is the reinforcement (or reward) signal provided by the environment at time t, and $e_i(t)$ is the eligibility trace for that connection. The eligibility trace $e_i(t)$ captures the recent history of activity on the ith connection and decays over time following

$$e_i(t+1) = \delta\, e_i(t) + (1 - \delta)\, y(t)\, x_i(t),$$

with δ representing the decay factor, $y(t)$ the output of the system at time t, and $x_i(t)$ the ith component of the input vector. Meanwhile, the critic's weight for the ith input is denoted by $v_i(t)$ and is updated by

$$v_i(t+1) = v_i(t) + \beta\, [r(t) + \gamma\, p(t) - p(t-1)]\, x_i(t),$$

where β is the critic's learning rate, γ is the discount factor that determines the influence of future rewards, and $p(t)$ is the critic's prediction of the cumulative future reward at time t. This architecture, by employing these update rules, addresses the credit-assignment problem by learning which actions under uncertain, delayed feedback contribute to success or failure.

3 Results

3.1 Data representation

We generated a synthetic dataset that simulates a minimally simplistic genomic variant data with both known and unknown pathogenicity. A necessary follow-up remains for the generation of highly accurate data representation and validation with real-world data . The dataset comprises key variables: GuRu Score, Population Frequency (*Pop Freq*), Gene Number (*Gene*), and ClinVar Pathogenicity (*Pathogenicity*). Known variants were produced from four distinct groups, reflecting varying profiles of these variables, while 50% of the dataset consists of variants with unknown pathogenicity, reserved for predictive modelling using reinforcement learning.

Figure 1 shows the overlaid distributions of these variables across three noise levels (0.1, 0.2, and 0.3). The figure illustrates how the distributions of *Guru Score*, *Pop Freq*, *Gene*, and *Pathogenicity* vary as the noise level increases, providing insight into the robustness of the generated data.

Figure 2 displays the correlation and covariance matrices for the known variants at each noise level. These matrices reveal notable relationships among the variables, such as a strong inverse correlation between *Guru Score* and *Pop Freq*, as well as a positive correlation between *Guru Score* and *Pathogenicity.* The covariance matrices further quantify the variability inherent in these relationships.

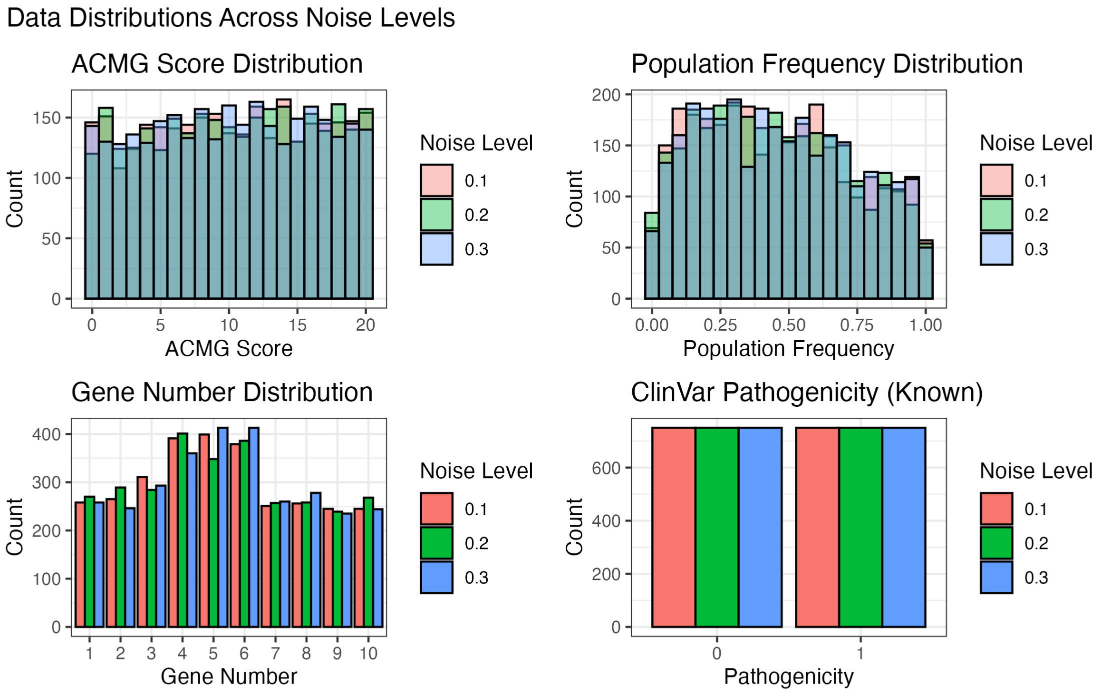
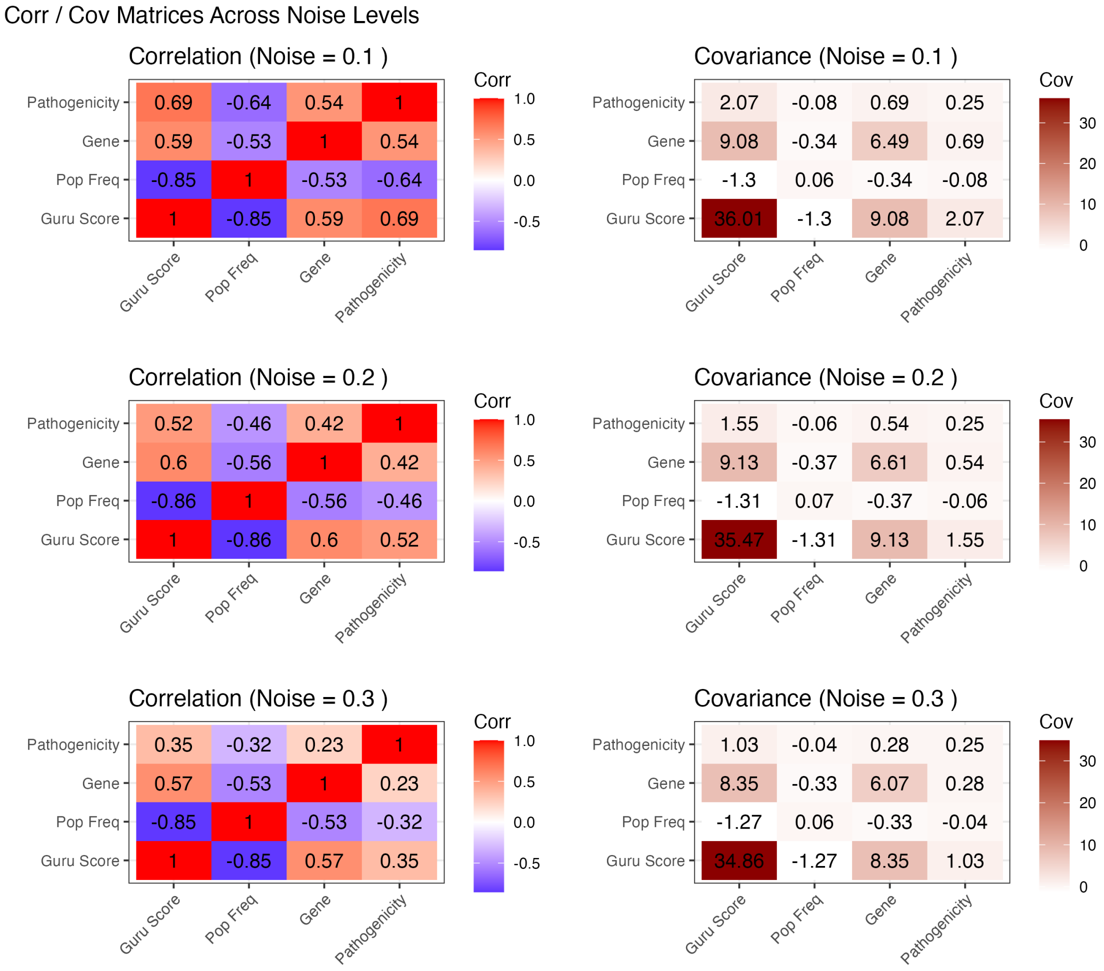


Figure 1: Data Distributions Across Noise Levels. This figure presents the overlaid distributions of GuRu Score, Population Frequency, Gene Number, and ClinVar Pathogenicity for noise levels 0.1, 0.2, and 0.3.



Correlation (Noise = 0.1)

	Guru Score	Pop Freq	Gene	Pathogenicity
Pathogenicity	0.69	-0.64	0.54	1
Gene	0.59	-0.53	1	0.54
Pop Freq	-0.85	1	-0.53	-0.64
Guru Score	1	-0.85	0.59	0.69

Covariance (Noise = 0.1)

	Guru Score	Pop Freq	Gene	Pathogenicity
Pathogenicity	2.07	-0.08	0.69	0.25
Gene	9.08	-0.34	6.49	0.69
Pop Freq	-1.3	0.06	-0.34	-0.08
Guru Score	36.01	-1.3	9.08	2.07

Correlation (Noise = 0.2)

	Guru Score	Pop Freq	Gene	Pathogenicity
Pathogenicity	0.52	-0.46	0.42	1
Gene	0.6	-0.56	1	0.42
Pop Freq	-0.86	1	-0.56	-0.46
Guru Score	1	-0.86	0.6	0.52

Covariance (Noise = 0.2)

	Guru Score	Pop Freq	Gene	Pathogenicity
Pathogenicity	1.55	-0.06	0.54	0.25
Gene	9.13	-0.37	6.61	0.54
Pop Freq	-1.31	0.07	-0.37	-0.06
Guru Score	35.47	-1.31	9.13	1.55

Correlation (Noise = 0.3)

	Guru Score	Pop Freq	Gene	Pathogenicity
Pathogenicity	0.35	-0.32	0.23	1
Gene	0.57	-0.53	1	0.23
Pop Freq	-0.85	1	-0.53	-0.32
Guru Score	1	-0.85	0.57	0.35

Covariance (Noise = 0.3)

	Guru Score	Pop Freq	Gene	Pathogenicity
Pathogenicity	1.03	-0.04	0.28	0.25
Gene	8.35	-0.33	6.07	0.28
Pop Freq	-1.27	0.06	-0.33	-0.04
Guru Score	34.86	-1.27	8.35	1.03

Figure 2: Correlation and Covariance Matrices Across Noise Levels. The figure displays the correlation (left) and covariance (right) matrices for the known variants, illustrating the relationships between *Guru Score*, *Pop Freq*, *Gene*, and *Pathogenicity* at various noise levels.

3.2 Performance Evaluation

In this study, we evaluated the performance of a RL model that uses GuRu scores and additional variant features to predict evidence supporting the assignment of pathogenicity in genetic data. The model was trained on known disease-causing variants and subsequently applied to predict the status of unknown variants. The experiments varied the noise level in the training labels, as well as the actor (α) and critic (β) learning rates. Performance was assessed using several metrics: average reward and TD error during training, cumulative learning curves, ROC curves with the corresponding AUC, and calibration of predicted probabilities.

Figure 3 shows the evolution of the average reward per epoch during training. This plot demonstrates that, under different parameter settings and noise levels, the model progressively improves its reward performance over successive epochs, indicating effective learning. Complementing this, Figure 4 presents the average TD error per epoch. The decreasing TD error over time reflects the convergence of the model's value estimates and confirms that the learning algorithm is stabilising across training epochs. Figure 5 depicts the cumulative average reward over training samples, providing a continuous view of the learning progress. The gradual increase in the cumulative reward further supports the model's ability to optimise its decision-making process over time.

For evaluating the classification performance, Figure 6 displays the ROC curves for various parameter combinations. These curves illustrate the trade-off between the true positive rate and the false positive rate, with several settings yielding discrimination ability. This observation is reinforced by the AUC heatmap in Figure 7, which summarises how AUC values vary as a function of α and β for each noise level. In some cases, AUC values exceed 0.8, demonstrating adequate baseline performance.

Figure 8 provides a calibration plot comparing the mean predicted probabilities to the observed proportions of pathogenic variants. Close alignment between the predicted and observed values for given parameter settings indicates that the model's probability estimates are well-calibrated. These results demonstrate that our baseline RL framework is capable of learning to predict evidence for assigning variant pathogenicity from noisy data, with the performance being sensitive to the learning rate parameters and the level of label noise.

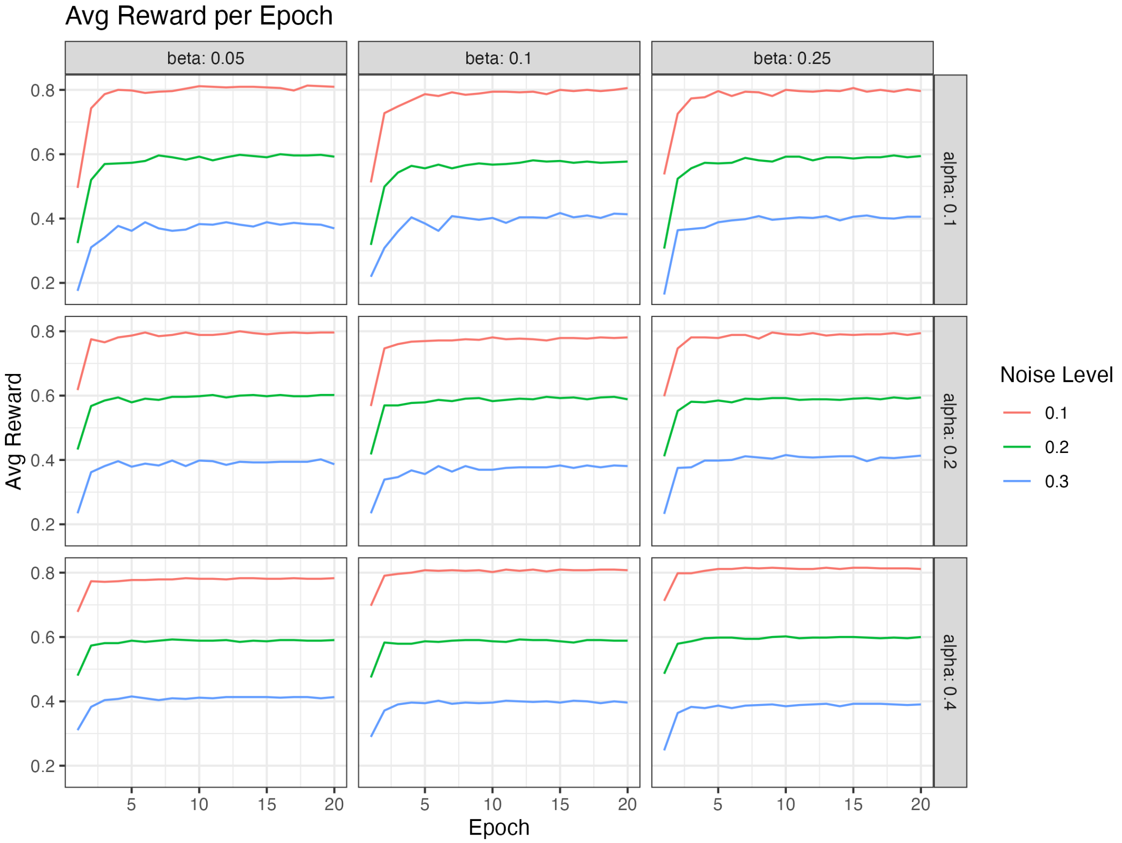


Figure 3: Average Reward per Epoch. This plot illustrates the evolution of the average reward during training, highlighting the model's improving performance across different parameter settings and noise levels.

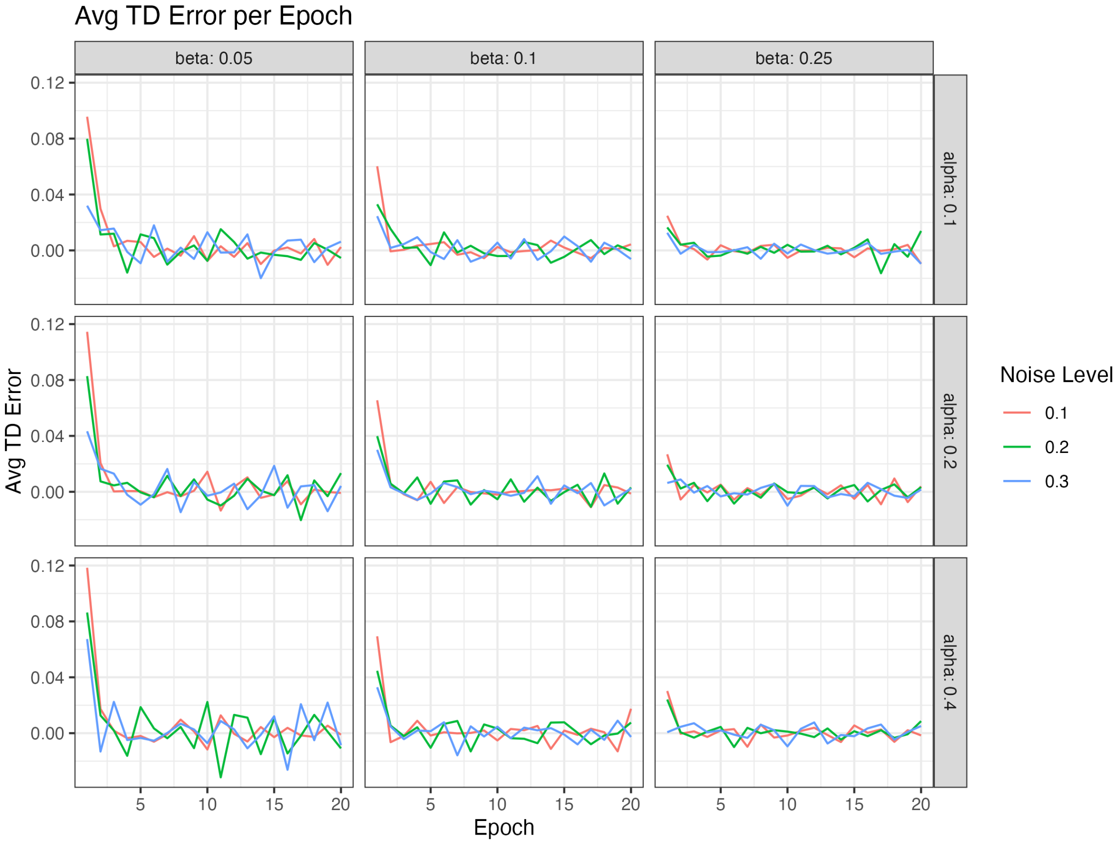


Figure 4: Average Temporal-Difference (TD) Error per Epoch. This figure displays the convergence behaviour of the learning algorithm, with decreasing TD error over successive epochs indicating stabilisation of the value estimates.

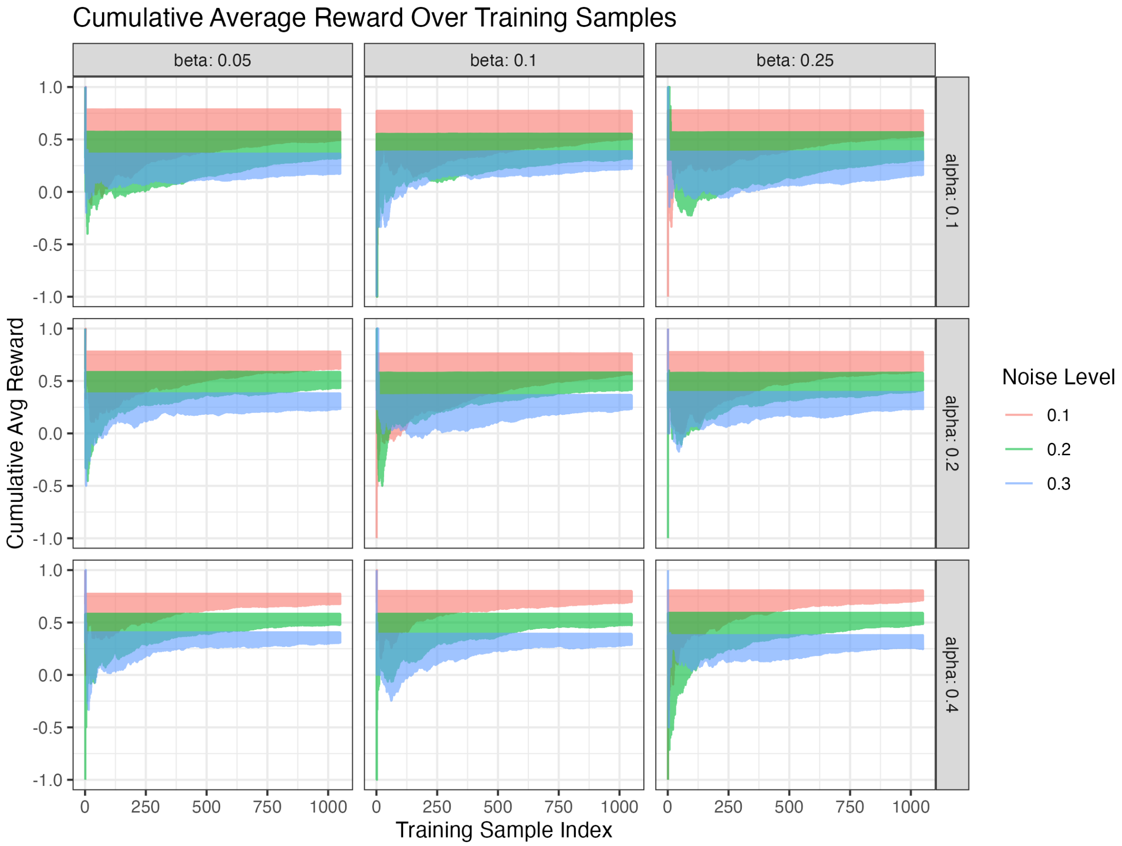


Figure 5: Cumulative Average Reward over Training Samples. The learning curve reflects the continuous improvement in model performance throughout training.

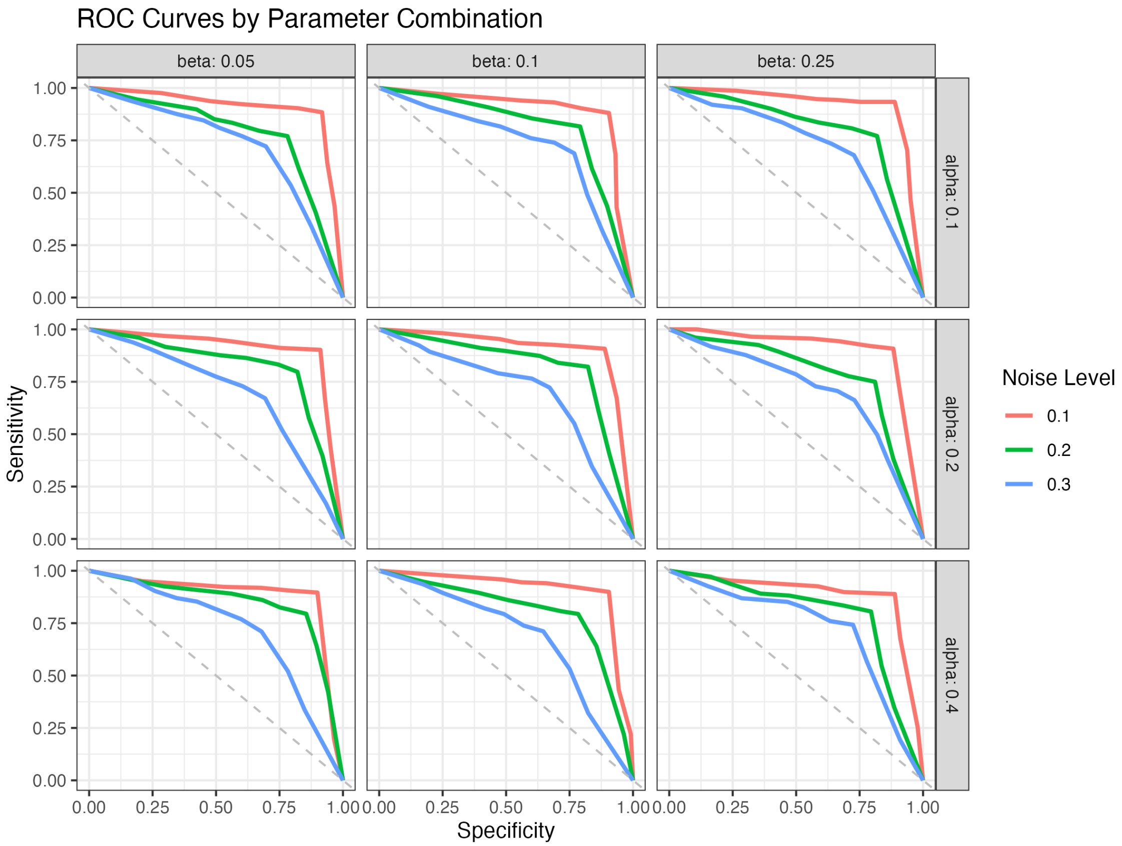


Figure 6: Receiver Operating Characteristic (ROC) Curves by Parameter Combination. These curves illustrate the trade-off between the true positive and false positive rates, demonstrating high discrimination ability under certain parameter settings.

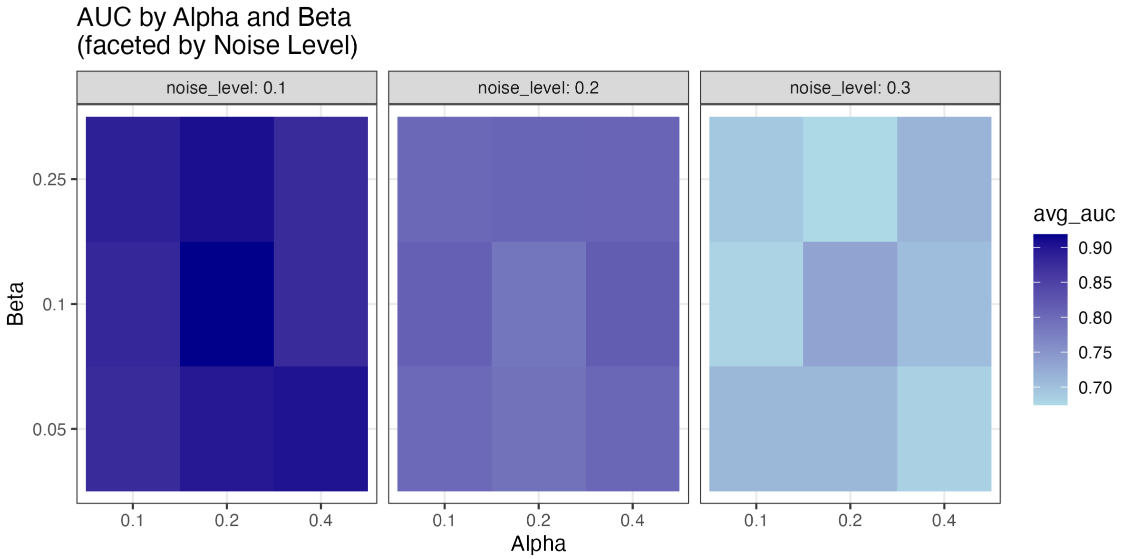


Figure 7: AUC Heatmap. This heatmap summarises the area under the ROC curve (AUC) as a function of the actor (α) and critic (β) learning rates, faceted by noise level. Several combinations achieve AUC values exceeding 0.8, indicating robust classification performance.

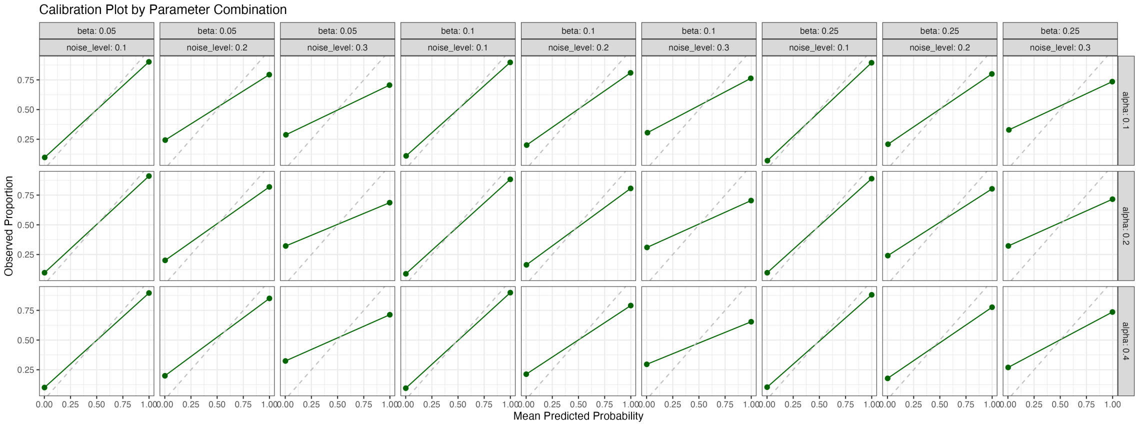


Figure 8: Calibration Plot. This plot compares the mean predicted probabilities to the observed proportions of pathogenic variants, showing that the predicted probabilities are well-calibrated across various parameter combinations.

4 Discussion

In this study, we have established an RL framework as an incremental step toward a broader Bayesian methodology aimed at classifying prior evidence about genetic variants being identified as pathogenic (or benign). Our current investigation focuses on evaluating RL methods on simulated data, explicitly quantifying multiple performance metrics relevant to the prediction of pathogenicity. By simulating scenarios reflective of real-world complexities, such as label noise and variant heterogeneity, we have demonstrated which baseline RL configurations exhibit most reliable performance.

The progressive improvement observed in average reward per epoch (Figure 3) illustrates the model's capacity for learning despite varying levels of noise and different hyperparameter settings. Notably, the reduction in TD error across epochs (Figure 4) indicates successful convergence of our actor–critic RL algorithm. The architecture is built on two key elements: an associative search element (ASE, the actor) and an adaptive critic element (ACE, the critic) (14). The ACE, which evaluates the consequences of actions (e.g. in pole-balancing) without accessing the actor's decisions, is adaptive-learning to predict long-term reward that, in turn, reinforces the actor's performance. Temporal-difference learning drives this process by using successive prediction differences to reduce errors and estimate future reward in a self-supervised manner (14). Early versions estimated the gradient of the objective function stochastically, later incorporating REINFORCE to achieve stochastic gradient ascent (albeit slower than backpropagation) (15–17).

Convergence analyses of actor-critic methods remain more complex than traditional policy iteration; comprehensive results demonstrate convergence to a local maximum using a two-timescale approach (with the critic learning faster than the actor) (18).

A critical aspect of reinforcement learning is its dynamic adaptability, which we have visualised in the cumulative average reward curves (Figure 5). This upward trajectory highlights the model's ability to progressively enhance predictive accuracy by addressing delayed reward problems-assigning appropriate credit or blame to actions long before their relevant outcomes appear. Such adaptability is imperative for practical genomic applications, where new sources of variant evidence continuously emerge, and datasets evolve.

Our analysis of ROC curves (Figure 6) and the corresponding heatmap of AUC values (Figure 7) provides a view of the predictive performance across varying parameters. Several parameter combinations produced robust discrimination performance, underscored by AUC values exceeding 0.8 even under challenging noise conditions. This finding emphasises the potential for selecting optimal RL hyperparameters tailored to specific genomic contexts and quality conditions in real datasets.

In addition to actor-critic algorithms, the 1990s witnessed the emergence of action-value methods, such as Q-learning, which map state–action pairs to expected returns and are sometimes referred to as "action-dependent adaptive critics" (19; 20). More recently, policy-gradient methods, including actor-critic algorithms, have gained traction due to their advantages in handling continuous action spaces, enabling probabilistic action selection that can converge to deterministic policies, and incorporating prior knowledge, potentially leading to faster and superior policies (21). The principles underlying these approaches have even influenced high-profile AI achievements, such as DeepMind's Go-playing programs, which, despite their complexity, share foundational elements like dynamic environment interaction, trial-and-error search, and long-term reward prediction (14; 22–24).

Calibration analysis (Figure 8) demonstrates that the RL model can generate accurate probability estimates that closely match observed pathogenic proportions. Reliable calibration is essential for clinical interpretation, as it provides confidence in the quantitative risk assessments derived from such models. The consistency observed across a variety of parameter settings strengthens the credibility of our RL-based predictions.

This study sets the stage for subsequent application of our RL framework to real genomic data, where the complexities encountered in simulations are amplified by biological variability and data imperfections. The current simulated environment has enabled us to systematically explore and quantify algorithm performance across controlled yet realistic scenarios. Our future work will focus on applying these RL methods to empirical genomic data, further integrating Bayesian frameworks to enhance interpretability, robustness, and clinical relevance.

Ultimately, this incremental methodological development has implications for broader genomic research. By refining predictive accuracy and interpretability, RL and subsequent Bayesian methodologies hold promise for transforming evidence interpretation, enhancing clinical decision-making, and ultimately improving personalised healthcare.

5 Conclusion

In this study, we have developed an actor-critic RL ramework that uses the GuRu score, variant/gene risk categorisation, and population frequency to estimate the probability of observing a variant in disease, rather than directly classifying its pathogenicity. A central contribution of our work is the demonstration that RL can learn to discern which evidence supports known variant labels, thereby serving as a crucial precursor to a broader Bayesian classification framework. Our evaluations indicate that the model achieves robust predictive performance even under varying noise conditions, as evidenced by reasonable baseline AUC values, well-calibrated probability estimates, and converging learning curves. These findings pave the way for future integration of RL with Bayesian inference in clinical genomics, ultimately enhancing the accuracy and genetic variant interpretation and improving personalized healthcare outcomes.

References

[1] Sue Richards, Nazneen Aziz, Sherri Bale, David Bick, Soma Das, Julie Gastier-Foster, Wayne W Grody, Madhuri Hegde, Elaine Lyon, Elaine Spector, 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. *Genetics in medicine*, 17 (5):405–423, 2015.

[2] Marilyn M Li, Michael Datto, Eric J Duncavage, Shashikant Kulkarni, Neal I Lindeman, Somak Roy, Apostolia M Tsimberidou, Cindy L Vnencak-Jones, Daynna J Wolff, Anas Younes, et al. Standards and guidelines for the interpretation and reporting of sequence variants in cancer: a joint consensus recommendation of the association for molecular pathology, american society of clinical oncology, and college of american pathologists. *The Journal of molecular diagnostics*, 19(1):4–23, 2017.

[3] Quan Li and Kai Wang. Intervar: clinical interpretation of genetic variants by the 2015 acmg-amp guidelines. *The American Journal of Human Genetics*, 100 (2):267–280, 2017.

[4] Erin Rooney Riggs, Erica F Andersen, Athena M Cherry, Sibel Kantarci, Hutton Kearney, Ankita Patel, Gordana Raca, Deborah I Ritter, Sarah T South, Erik C Thorland, et al. Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the american college of medical genetics and genomics (acmge and the clinical genome resource (clingen). *Genetics in Medicine*, 22(2):245–257, 2020.

[5] Sean V Tavtigian, Steven M Harrison, Kenneth M Boucher, and Leslie G Biesecker. Fitting a naturally scaled point system to the acmg/amp variant classification guidelines. *Human mutation*, 41(10):1734–1737, 2020.

[6] Brent S Pedersen, Joe M Brown, Harriet Dashnow, Amelia D Wallace, Matt Velinder, Martin Tristani-Firouzi, Joshua D Schiffman, Tatiana Tvrdik, Rong Mao, D Hunter Best, et al. Effective variant filtering and expected candidate variant yield in studies of rare human disease. *NPJ Genomic Medicine*, 6(1):1–8, 2021.

[7] Carl A Anderson, Fredrik H Pettersson, Geraldine M Clarke, Lon R Cardon, Andrew P Morris, and Krina T Zondervan. Data quality control in genetic

case-control association studies. *Nature protocols*, 5(9):1564–1573, 2010. URL https://doi.org/10.1038/nprot.2010.116.

[8] Emil Uffelmann, Qin Qin Huang, Nchangwi Syntia Munung, Jantina De Vries, Yukinori Okada, Alicia R Martin, Hilary C Martin, Tuuli Lappalainen, and Danielle Posthuma. Genome-wide association studies. *Nature Reviews Methods Primers*, 1(1):59, 2021. URL https://doi.org/10.1038/s43586-021-00056-9.

[9] Konrad J Karczewski, Laurent C Francioli, Grace Tiao, Beryl B Cummings, Jessica Alföldi, Qingbo Wang, Ryan L Collins, Kristen M Laricchia, Andrea Ganna, Daniel P Birnbaum, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581(7809):434–443, 2020.

[10] Melissa J Landrum, Jennifer M Lee, Mark Benson, Garth Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Jeffrey Hoover, et al. Clinvar: public archive of interpretations of clinically relevant variants. *Nucleic acids research*, 44(D1):D862–D868, 2016.

[11] Edgar A Rivera-Muñoz, Laura V Milko, Steven M Harrison, Danielle R Azzariti, C Lisa Kurtz, Kristy Lee, Jessica L Mester, Meredith A Weaver, Erin Currey, William Craigen, et al. Clingen variant curation expert panel experiences and standardized processes for disease and gene-level specification of the acmg/amp guidelines for sequence variant interpretation. *Human mutation*, 39(11):1614–1622, 2018.

[12] Andrew G Barto, Richard S Sutton, and Charles W Anderson. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE transactions on systems, man, and cybernetics*, (5):834–846, 1983.

[13] Andrew G Barto, Richard S Sutton, and Charles W Anderson. Looking back on the actor–critic architecture. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 51(1):40–50, 2020.

[14] Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

[15] David E Rumelhart, Geoffrey E Hinton, Ronald J Williams, et al. Learning internal representations by error propagation, 1985.

[16] Andrew G Barto and Michael I Jordan. Gradient following without backpropagation in layered networks. et-al. *Frontiers in cognitive neuroscience*, pages 443–449, 1992.

[17] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256, 1992.

[18] Shalabh Bhatnagar, Richard S Sutton, Mohammad Ghavamzadeh, and Mark Lee. Naturalgradient actor-critic algorithms. *Automatica*, 2007.

[19] Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8: 279–292, 1992.

[20] Derong Liu, Xiaoxu Xiong, and Yi Zhang. Action-dependent adaptive critic designs. In *IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No. 01CH37222)*, volume 2, pages 990–995. IEEE, 2001.

[21] Charles W Anderson. Approximating a policy can be easier than approximating a value function. *Computer Science Technical Report*, 2000.

[22] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

[23] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

[24] Marco A Wiering and Martijn Van Otterlo. Reinforcement learning. *Adaptation, learning, and optimization*, 12(3):729, 2012.