An Actor-Critic Reinforcement Learning Framework for Genetic Variant Pathogenicity

Dylan Lawless*[1]

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

March 13, 2025

Acronyms

Abstract

We present a reinforcement learning (RL) framework designed to predict genetic variant pathogenicity by integrating established metrics such as the American College of Medical Genetics and Genomics (ACMG) score, gene risk factors, and population frequency. Using simulated genetic data with realistic variability and label noise, we systematically evaluated RL performance across multiple scenarios, employing metrics including ROC curves, AUC, calibration plots, and learning dynamics. Results indicate robust 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.

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

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.

Reinforcement learning (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 variant pathogenicity.

In this study, we systematically quantify RL performance using a comprehensive suite of evaluation metrics within simulated genomic scenarios, incorporating realistic data imperfections. Our goal is to identify RL methodologies and parameter configurations that exhibit robust and accurate predictive capabilities, thus laying the groundwork for subsequent integration into a Bayesian analytical framework aimed at interpreting genetic evidence in clinical diagnostics.

2 Methods

Our investigation employs a synthetic dataset designed to mimic the characteristics of genetic variant data. The dataset comprises 10,000 variants, with 50% of the entries containing known pathogenicity labels and the remaining 50% left unlabelled for prediction purposes. Variants are generated to reflect realistic genomic scenarios by incorporating features such as the GuRu score, gene number, and population frequency. Notably, variants occurring in genes numbered 4 to 10 are considered to be of higher risk, while those in genes 1 to 6 are deemed to be of lower risk. Furthermore, to replicate potential misannotations in empirical data, a proportion of the known labels is deliberately flipped according to a predefined noise parameter.

The reinforcement learning framework is implemented via an actor-critic algorithm. In our formulation, the state space is 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 yields 24 unique states. The action space is binary, corresponding to the two possible classifications: benign (0) or pathogenic (1). At each iteration, the RL agent observes a state corresponding to a variant and selects an action using a probabilistic policy derived from a sigmoid function applied to a vector of actor weights. A reward of +1 is granted if the predicted label matches the true pathogenicity, otherwise a penalty of -1 is imposed. The temporal-difference (TD) error, defined as the difference between the received reward and the critic's estimate of the state value, is 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 vary three critical parameters: the noise level in the training labels, the actor learning rate, and the critic learning rate. For each combination of these parameters, the model is trained for 20 epochs and a variety of performance metrics are recorded. The evaluation comprises 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 are aggregated and visualised using faceted plots, thus allowing for a comprehensive comparison of the effects of different parameter settings.

The complete experimental implementation is written in R. Synthetic data are generated according to the specified stratification and noise levels, and the RL model is trained on a randomly partitioned training set for each combination of parameters. Performance on a held-out test set is evaluated and visualised through several plots, each of which is saved as an individual PNG file for further examination.

3 Results

In this study, we evaluated the performance of a reinforcement learning model that leverages GuRu scores and additional variant features to predict 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 temporal-difference (TD) error during training, cumulative learning curves, receiver operating characteristic (ROC) curves with the corresponding area under the curve (AUC), and calibration of predicted probabilities.

Figure 1 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 2 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 3 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 4 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 high discrimination ability. This observation is reinforced by the AUC heatmap in Figure 5, 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 robust performance.

Finally, Figure 6 provides a calibration plot comparing the mean predicted probabilities to the observed proportions of pathogenic variants. The close alignment between the predicted and observed values across different parameter settings indicates that the model's probability estimates are well-calibrated.

Overall, these results demonstrate that our reinforcement learning framework is capable of effectively learning to predict variant pathogenicity from noisy data, with the performance being sensitive to the learning rate parameters and the level of label noise.

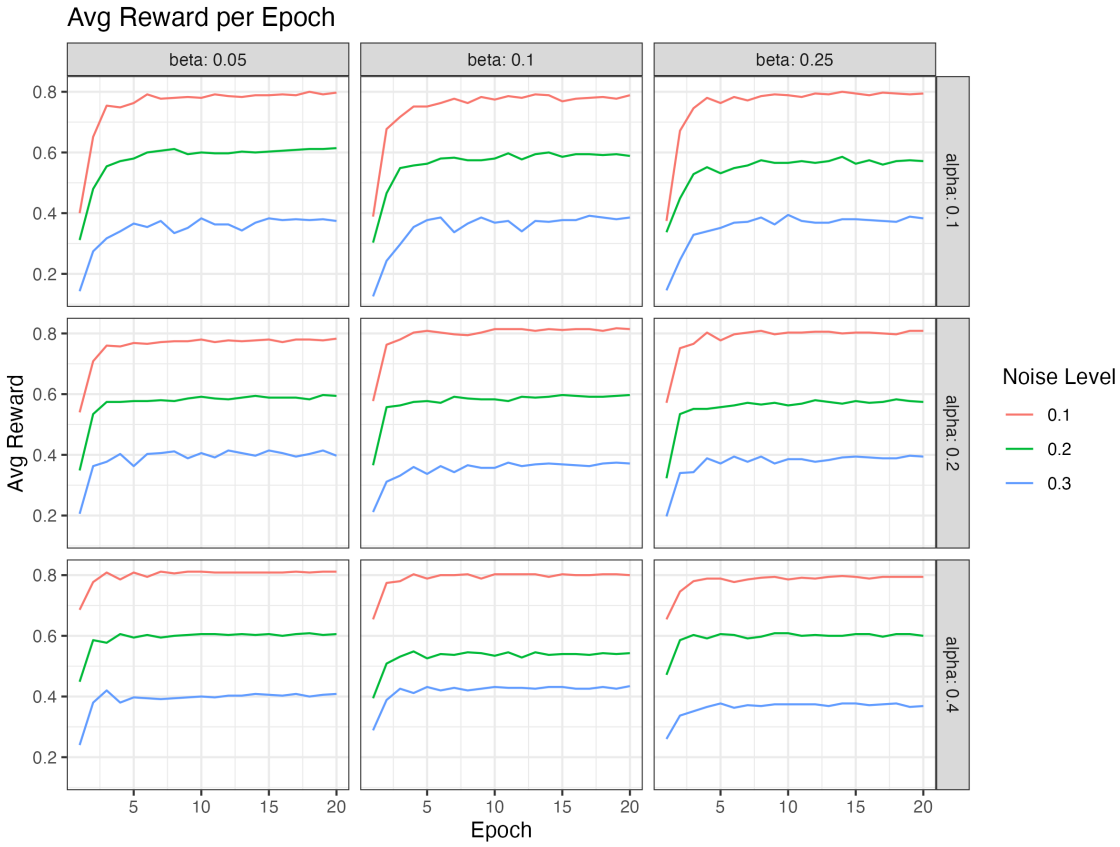

Figure 1: Average Reward per Epoch. This plot illustrates the evolution of the average reward during training, highlighting the model's improving performance across different parameter settings and noise levels.

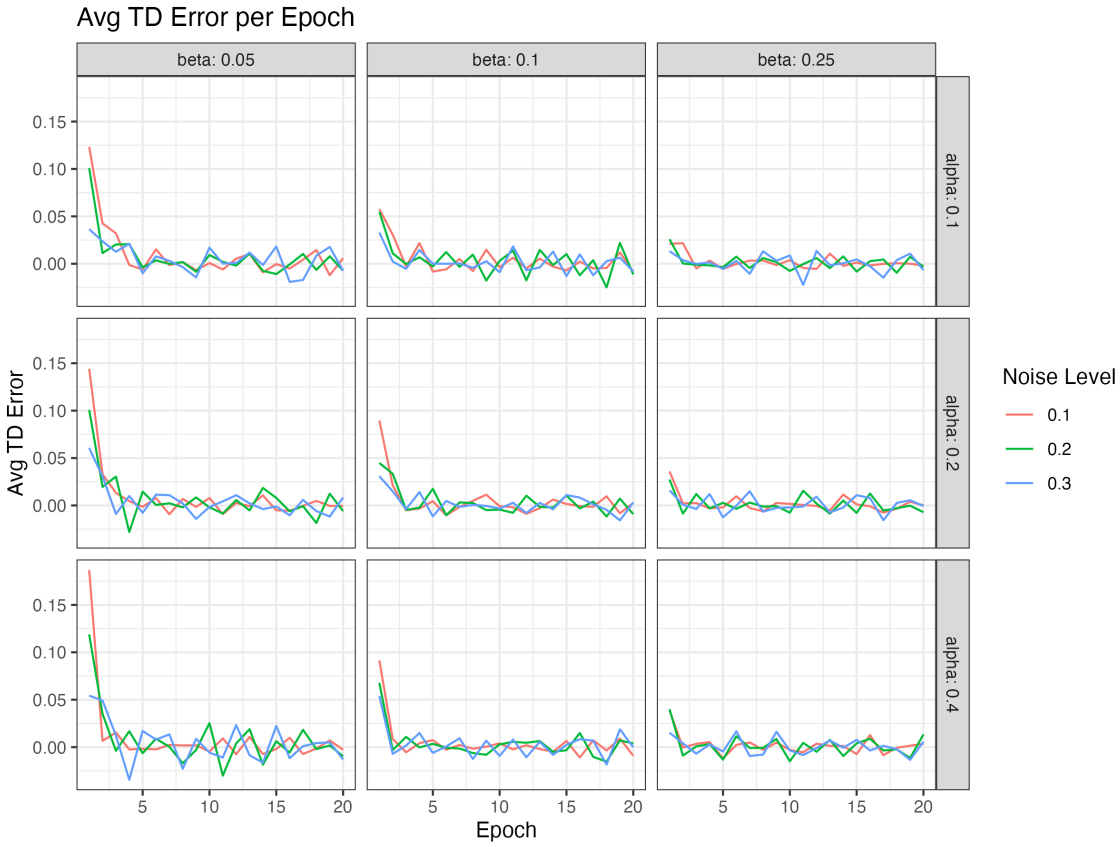

Figure 2: Average Temporal-Difference (TD) Error per Epoch. This figure displays the convergence behaviour of the learning algorithm, with decreasing TD error over successive epochs indicating stabilisation of the value estimates.

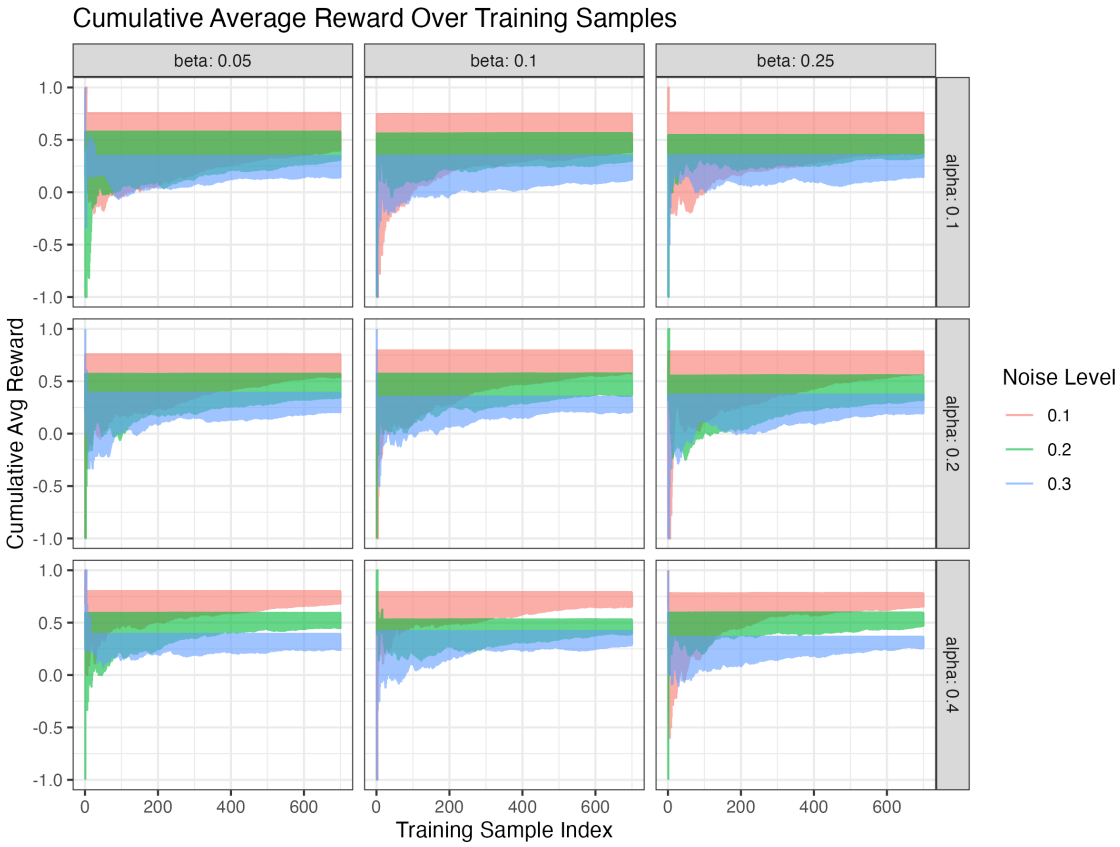

Figure 3: Cumulative Average Reward over Training Samples. The learning curve reflects the continuous improvement in model performance throughout training.

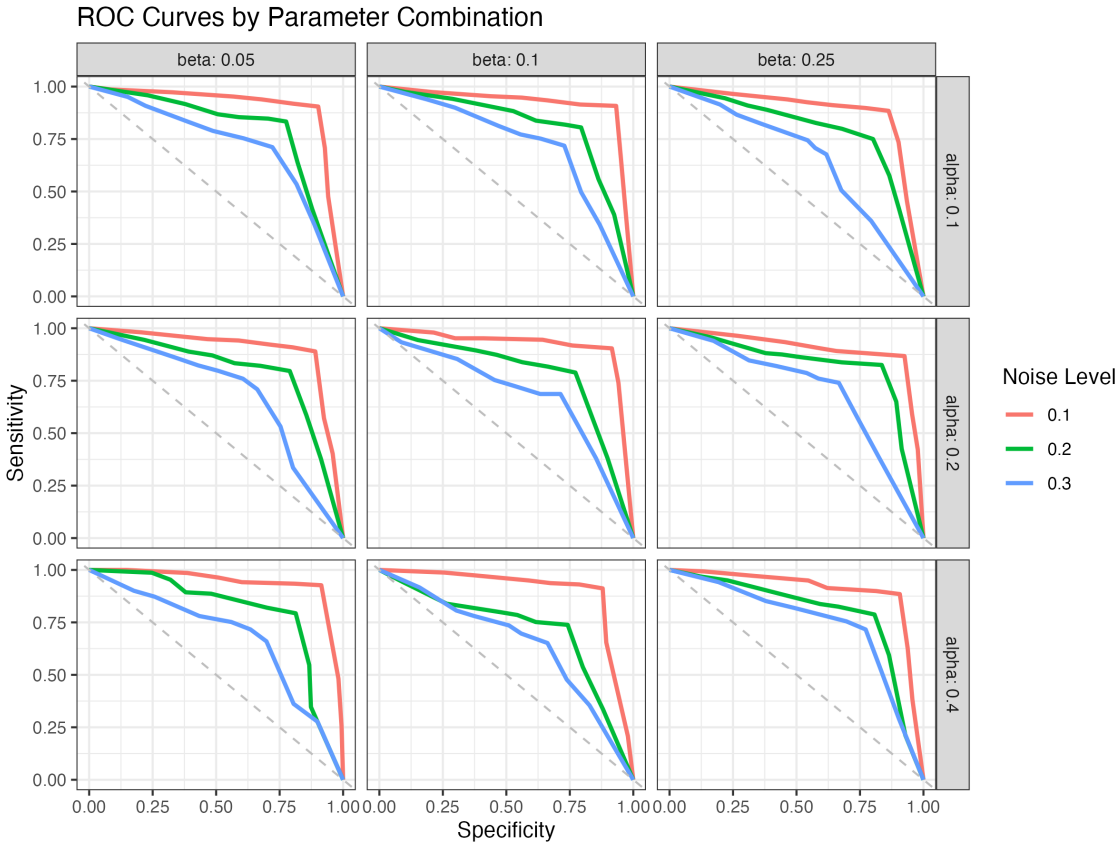

Figure 4: Receiver Operating Characteristic (ROC) Curves by Parameter Combination. These curves illustrate the trade-off between the true positive and false positive rates, demonstrating high discrimination ability under certain parameter settings.

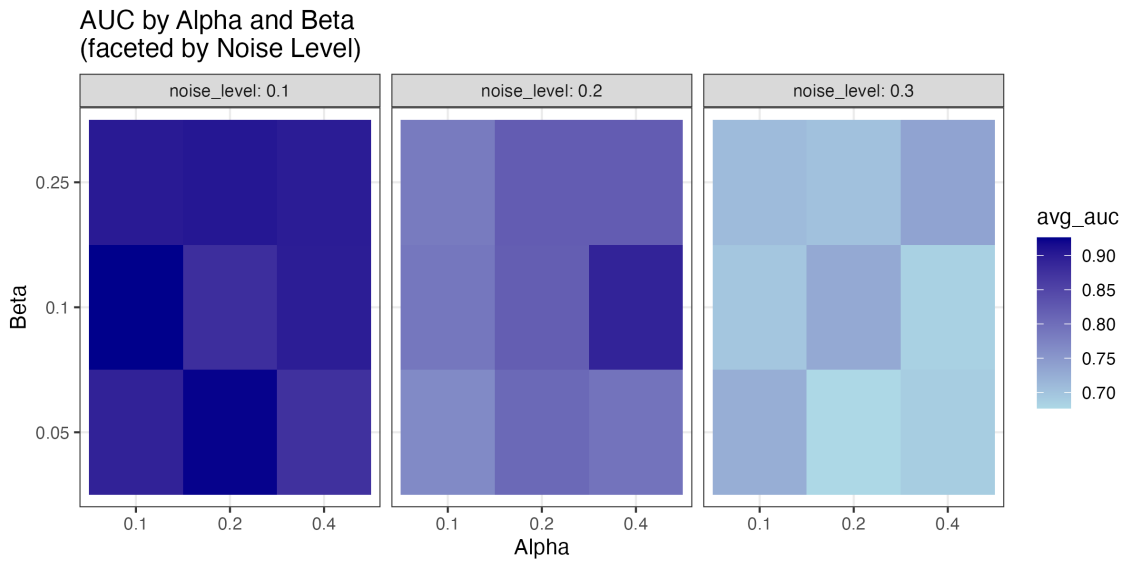

Figure 5: AUC Heatmap. This heatmap summarises the area under the ROC curve (AUC) as a function of the actor (α) and critic (β) learning rates, faceted by noise level. Several combinations achieve AUC values exceeding 0.8, indicating robust classification performance.

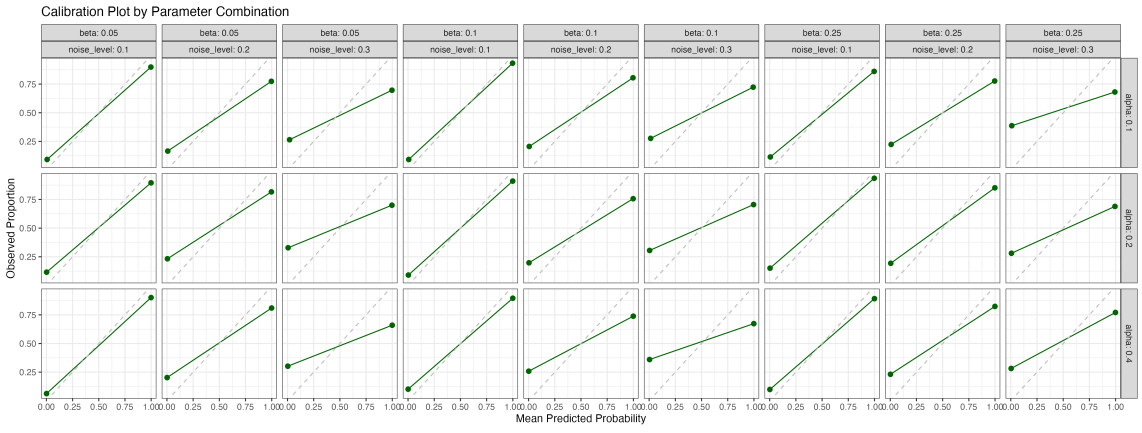

Figure 6: Calibration Plot. This plot compares the mean predicted probabilities to the observed proportions of pathogenic variants, showing that the predicted probabilities are well-calibrated across various parameter combinations.

4 Discussion

In this study, we have established a reinforcement learning (RL) framework as an incremental step toward a broader Bayesian methodology aimed at classifying genetic variants for their pathogenic potential. 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 elucidated which RL configurations exhibit robust and reliable performance.

The progressive improvement observed in average reward per epoch (Figure 1) illustrates the model's capacity for effective learning despite varying levels of noise and different hyperparameter settings. Notably, the reduction in temporal-difference (TD) error across epochs (Figure 2) indicates successful convergence of our actor-critic RL algorithm. This stabilisation is critical, as it demonstrates the algorithm's ability to adaptively refine its predictions, which is a cornerstone for its utility in clinical genomic interpretation.

A critical aspect of reinforcement learning is its dynamic adaptability, which we have visualised clearly in the cumulative average reward curves (Figure 3). This continuous upward trajectory highlights the model's ability to enhance predictive accuracy progressively throughout the training period. Such adaptability will be indispensable in practical genomic applications, where new variants continuously emerge, and datasets evolve.

Our analysis of ROC curves (Figure 4) and the corresponding heatmap of AUC values (Figure 5) provides a comprehensive 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.

Calibration analysis (Figure 6) 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 significant implications for the broader genomic research community. By refining predictive accuracy and interpretability, RL and subsequent Bayesian methodologies hold promise for transforming genetic diagnostics, enhancing clinical decision-making, and ultimately improving personalised healthcare.

5 Conclusion

We have presented an actor-critic reinforcement learning framework that leverages the GuRu score, gene number, and population frequency to predict the pathogenicity of genetic variants. By training on a dataset of known disease-causing variants and evaluating on unlabelled data, the model demonstrates promising performance across a range of parameter settings. The comprehensive evaluation, which includes ROC analysis, precision-recall metrics, learning curves and calibration assessments, highlights the potential of RL methodologies in the field of genetic diagnostics. This framework provides a solid foundation for further research into the integration of machine learning techniques with genomic data analysis and may ultimately contribute to improved personalised medical interventions.