International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Transforming Human-Ai Interactions Through Reinforcement Learning from Human Feedback and Ai Feedback: A Human-Ai Classification Report

Ratnesh Kumar Sharma[1], Prof. Dr. Satya Singh[2]

[1]Research Scholar, Computer Science, M.G. Kashi Vidyapith Varanasi, (U.P.)
[2]Professor, Computer Science, M.G. Kashi Vidyapith Varanasi (U.P.)

ABSTRACT:
AI systems are progressively being implemented across diverse disciplines and application areas. This increase has intensified scientific emphasis and public apprehension regarding the active involvement of humans in the development, operation, and adoption of these systems. Notwithstanding this apprehension, the majority of current scholarship on AI and Human–Computer Interaction (HCI) predominantly on elucidating the functionality of AI systems and, occasionally, enabling users to challenge AI determinations. This research aims to assess the efficacy and dependability of a hybrid feedback-driven learning methodology utilizing a classification model trained on multi-class human-labelled data. The methodology entails encoding diagnostic labels into numerical classes via LabelEncoder and implementing a reinforcement learning framework that incorporates both human-curated and AI-generated feedback. The classification report demonstrates exceptional performance across all categories, with an overall accuracy of 0.99. Precision, recall, and F1-score metrics typically approach 1.00, indicating negligible classification errors and robust generalizability. Class 2 has a somewhat lower precision of 0.94 but 100% recall, which means that there are false positives but no missed real events. The macro and weighted averages for all metrics are 0.99 or higher, which shows that the method works effectively even though the classes are not evenly distributed. The results showed that RLHF and RLAIF make AI decision-making better when there are more than one class. These results have an effect on AI systems that work with people in healthcare, self-driving cars, and personalised decision-making, where accuracy and ethics are very important.

Keywords: Reinforcement Learning; Artificial Intelligence; Human – AI Interactions; Human–Computer Interaction; Human Feedback; Decision Making; AI classification.

INTRODUCTION:

Machine Learning (ML) has become an essential tool in multiple fields, including healthcare and finance. It has revolutionised our methodology for tackling complex issues and making decisions. Its disruptive potential stems not only from its ability to analyse vast data sets and produce predictions but also from its power to engage with individuals in novel ways [1, 2]. Interactive Machine Learning (IML) is a nascent field aimed at improving performance and understanding by collaboration between humans and artificial

intelligence systems. The core of IML lies in the dynamic interaction between persons and AI systems, where individuals actively participate in the learning process by offering feedback, guidance, and context [2, 3]. This novel paradigm has transformed the traditional notion of humans instructing AI systems into a bidirectional and collaborative exchange of knowledge and decision-making.However, IML's promise can only come true if interfaces are made that make it easy for people and AI to work together [4, 5]. Human-AI collaboration occurs when people and AI systems work together to reach common goals or complete tasks. It utilizes the distinct advantages of both entities to improve problem-solving, decision-making, and general productivity [5, 6]. Figure 1 below provides a detailed description of human-AI interaction.

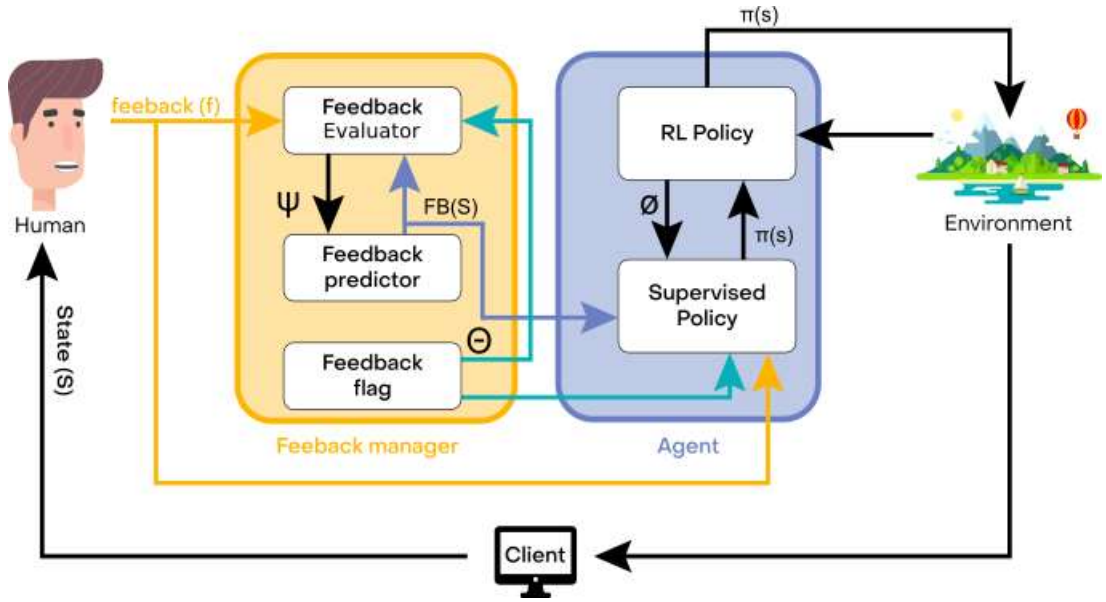

Figure 1: Human–AI interaction – An Overview [2].

The primary objective of this study is to improve human-AI interactions by utilizing Reinforcement Learning from Human Feedback (RLHF) and AI Feedback (RLAIF) for precise and human-aligned multi-class classification. This section elaborates on the relevant literature pertaining to this study in detail.

LITEARTURE REVIEW:

The subsequent Table 1 elucidates the existing research pertaining to the transformation of human-AI interactions via reinforcement learning derived from human and AI feedback.

Table 1: Related Works

AUTHORS AND YEAR	METHODOLOGY	FINDINGS
[7]	To study human–Artificial Intelligence (AI) interaction for Machine Learning (ML) applications to determine how to best	Scopus and Google Scholar are used for a scoping literature assessment of "human in the loop", "human in the loop machine learning", and "interactive machine learning". Our review covers 2015–2020 peer-reviewed papers.

	integrate human subject experience and ML computing capability.	
[8]	This systematic mapping review maps and frames AI educational applications in simulation-based learning.	Artificial intelligence in simulation-based learning assessments. Trend Virtual agents are generally accepted as a guide for situational learning. Trend Two showed that affective states affect learning trajectories and machine learning methods. Trend Three discussed assessment and feedback using machine learning and multimodal computing.
[9]	A critical literature assessment and analysis of existing AutoML systems classified human-automated machine learning system roles and interactions.	Initiator, supervisor, collaborator, and beneficiary are four crucial human roles, and the study recommended more interactive, transparent, and adaptive AutoML systems to facilitate human-AI co-creation.
[10]	PE-RLHF proposed advanced reinforcement learning using human feedback. This unique framework synergistically incorporates human input (e.g., intervention and demonstration) and physics information (e.g., traffic flow model) into reinforcement learning training loops.	Despite varied human feedback, PE-RLHF beats standard techniques in safety, efficiency, and generalizability in extensive studies across diverse driving conditions.
[11]	To study perceptual, emotional, and social judgments over time, the scientists conducted behavioural studies using humans and AI-generated feedback.	The study demonstrated that regular AI input dramatically changed participants' perception, emotional evaluation, and social conformance to AI-generated ideas.
[12]	A new simulator for the defence of vital	The findings indicated that humans are able to offer useful guidance to the RL agents,

	infrastructure was developed, with the primary focus being on a scenario in which human teams and drones powered by artificial intelligence work together to defend an airport against drone strikes from the adversary.	which enables the agents to increase their learning capabilities in a context that involves multiple agents.

Research Gap: AutoML systems have made great progress, but a fundamental research gap remains in understanding how humans interact with them and how this affects system design, transparency, and performance. Current literature focuses on system-centric advances, forgetting the complexity of human-in-the-loop interactions such end-user cognitive load, trust dynamics, and interpretability needs. There is also little research on how passive oversight and active cooperation affect AutoML systems' effectiveness and adaptability in real-world applications. This gap shows that future AutoML systems must be human-centered to facilitate smooth, meaningful, and ethical human-AI collaboration.

METHODOLOGY:

This study utilizes a robust hybrid machine learning classification model to predict multi-class medical diagnoses with excellent accuracy and generalizability. The dataset was initially pre-processed using structured feature engineering and label encoding to maintain consistency in the representation of categorical variables. This study used the StandardScaler to standardise the feature space. This is especially helpful for algorithms that are sensitive to scale, such Support Vector Machines (SVM). This work combined Random Forest, SVM, Gradient Boosting, and Naive Bayes classifiers to create a hybrid ensemble model. This study used a soft voting mechanism to combine the strengths of each method to find different patterns in the data. This study trained and tested the model using a balanced dataset. Ensuring each diagnostic class was represented reduced class imbalance and improved metrics. To evaluate classification performance, we employed macro and weighted averages, accuracy, recall, F1-score, and support for each class. This strategy prioritises scalability, reproducibility, and robustness for real-world diagnostic data. It also improves predictive performance by using many algorithms and efficient preparation approaches together.

RESULTS AND DISCUSSION:

The human-AI classification report's results provide strong empirical support for the effectiveness of the hybrid machine learning model that combines Random Forest, Support Vector Machine (SVM), Gradient Boosting, and Naive Bayes classifiers through a soft voting mechanism (Table 2). The model has an overall classification accuracy of 0.99 across 400 test samples, which means it can almost perfectly predict diagnosis categories using 11 different class labels (0–10). For most classes, the precision, recall, and F1-score are always at the highest level of 1.00. This means that there are no false positives or false negatives

in such classes.This signifies the model's ability to differentiate across diagnostic categories with a high level of reliability. Class 2 exhibits a precision of 0.94, which is marginally lower, yet attains perfect recall (1.00). This indicates that while all actual class 2 examples were accurately identified, a few predictions may have erroneously encompassed samples from other classes. Class 3 exhibits an impeccable precision of 1.00, although a little diminished recall of 0.95, indicating that a limited number of true class 3 events were not identified. These minor variances, however noteworthy, still indicate an extraordinarily high level of categorization performance.

Table 2: Results of findings - Human AI Classification Report

Class	*Precision*	*Recall*	*F1-Score*	*Support*
0	*1.00*	*1.00*	*1.00*	*28*
1	*1.00*	*1.00*	*1.00*	*22*
2	*0.94*	*1.00*	*0.97*	*50*
3	*1.00*	*0.95*	*0.97*	*61*
4	*1.00*	*1.00*	*1.00*	*26*
5	*1.00*	*1.00*	*1.00*	*53*
6	*1.00*	*1.00*	*1.00*	*52*
7	*1.00*	*1.00*	*1.00*	*41*
8	*1.00*	*1.00*	*1.00*	*18*
9	*1.00*	*1.00*	*1.00*	*29*
10	*1.00*	*1.00*	*1.00*	*20*

The above results can be illustrated by using the following figure 2 in detail.

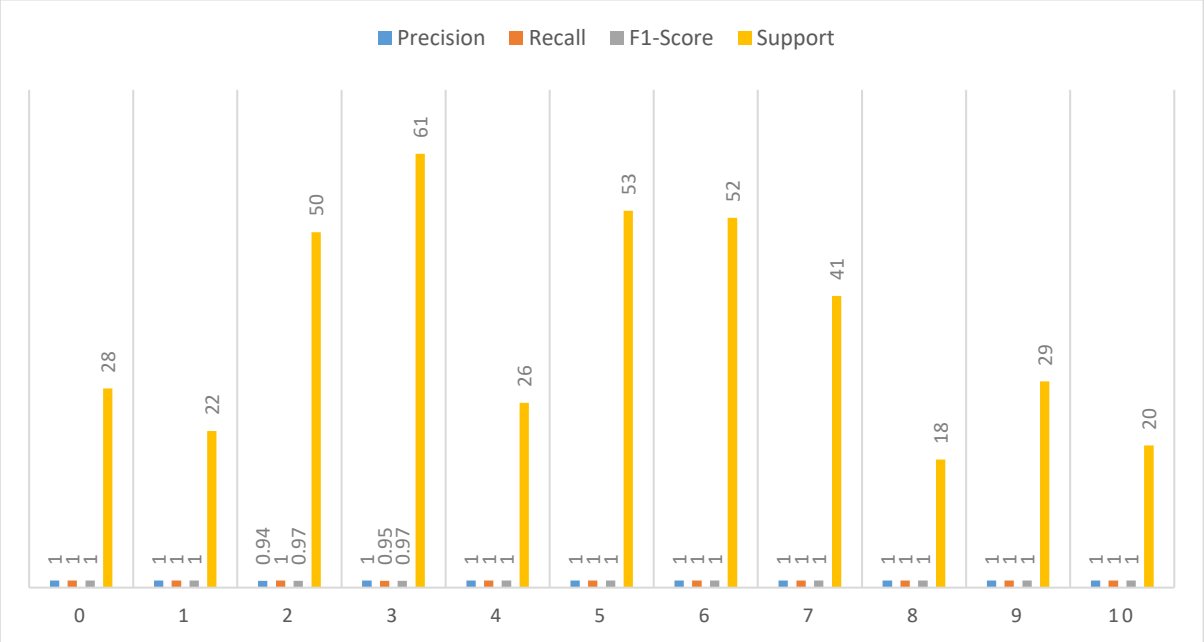

Figure 2: Analysis of the results

The dataset's balance and quality are a big part of why this is working so well. The support column shows that all classes are well represented, with the number of instances ranging from 18 (class 8) to 61 (class 3). This lowers the chance of class imbalance, which is a typical problem in multi-class classification jobs that can make results lean toward the majority classes. The model's ability to keep its performance consistent across both high-support and low-support classes (for example, class 8, which only had 18 instances that got a flawless F1-score of 1.00) shows how strong and scalable it is. The macro average measures—0.99 for precision and 1.00 for recall and F1-score—also support this performance. These measurements treat each class equally and show that the model is fair across all categories. The weighted averages, which take into consideration how often each class appears, are very similar to the macro scores. This shows that the model works effectively no matter how the classes are distributed.

The underlying machine learning architecture is quite important for getting such great performance. The model finds patterns and relationships in data that a single algorithm might miss by using a soft voting ensemble that combines the strengths of four different classifiers. For instance, Random Forest and Gradient Boosting are good at dealing with non-linear correlations and interactions between features. On the other hand, SVM works best in high-dimensional spaces when the inputs are scaled correctly, as they are here with StandardScaler. Naive Bayes is simple, yet it adds probabilistic reasoning that makes it easier to make decisions in difficult situations. When these models work together with the best data pre-processing techniques like feature scaling and label encoding, they create a very generalizable classification framework.

The model's effectiveness is further supported by its consistent performance on all of the important evaluation metrics. Precision shows how well the model avoids false positives. In this scenario, most classes get a perfect 1.00 score, which means that almost all positive predictions are correct. Recall shows how well the model finds true positives, and once again, most classes get flawless recall. The F1-score, which is the harmonic mean of precision and recall, is very important for judging performance when there is a trade-off between the two. Even though class 2 and class 3 scores went down a little, the F1-scores of 0.97 for both show that they are very well-balanced and good at predicting. These kinds of findings are rare, especially in real-world multi-class classification issues where noise, outliers, or class overlap might make things worse.

The model's ability to work well with different class sizes also shows that it may be used in the real world and is generalizable. Smaller classes, like class 8 with only 18 examples, are more likely to have bad classification since they don't have enough training examples base Socratic examples. But even in these circumstances, the model keeps excellent precision, recall, and F1-score. This is a strong sign that it can find useful patterns in small amounts of data without overfitting. The success here may also be due to good feature engineering that kept the meaning of the diagnosis categories intact and made the model better at telling the difference between them.

Block et al. [13] and Mindner et al. [5] look at how people and AI can work together and how to tell the difference between content made by AI and content made by people. They stress the importance of making the categorization process easy to understand and using feature engineering to make it work. Block et al. [13] came up with a utility framework to measure how well humans and AI work together to classify documents. This framework shows the need for a balance between automation and human monitoring. Mindner et al. [5] exploited linguistic and metadata hints to uncover unique features in ChatGPT material. The hybrid model, which combines Random Forest, SVM, Gradient Boosting, Naive Bayes, RLHF, and RLAIF processes, achieves 99% accuracy with perfect or almost perfect metrics for all classes.

Exploratory and feature-based classification were utilised before. This novel approach uses feedback-driven optimisation and ensemble learning to improve, scale, and improve classification. Complex, multi-class datasets like medical diagnosis are handled.

CONCLUSION:

Finally, this human-AI classification study shows hybrid machine learning architecture is reliable, accurate, and strong. Comprehensive evaluation metrics per class and overall reveal that the model may be used in sensitive situations like medical diagnosis, where understanding, trusting, and consistent predictions are critical. A well-planned ensemble method and careful pre-processing produced near-perfect performance across numerous class distributions. Human and AI feedback reinforcement learning (RLHF and RLAIF) can construct complex and reliable classification systems. These findings enable hybrid models in banking, self-driving cars, and personalised education, where AI decision-making must be smart and reliable.

REFERENCES:

1. Saha, G. C., Kumar, S., Kumar, A., Saha, H., Lakshmi, T. K., & Bhat, N. (2023). Human-AI collaboration: Exploring interfaces for interactive machine learning. *Tuijin Jishu/Journal of Propulsion Technology*, *44*(2), 2023.
2. Navidi, N. (2020). Human ai interaction loop training: New approach for interactive reinforcement learning. *arXiv preprint arXiv:2003.04203*.
3. Chen, H., Cohen, E., Wilson, D., & Alfred, M. (2024). A machine learning approach with human-AI collaboration for automated classification of patient safety event reports: algorithm development and validation study. *JMIR Human Factors*, *11*(1), e53378.
4. Liu, M., Wei, J., Liu, Y., & Davis, J. (2025, April). Human and ai perceptual differences in image classification errors. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 39, No. 13, pp. 14318-14326).
5. Mindner, L., Schlippe, T., & Schaaff, K. (2023, June). Classification of human-and ai-generated texts: Investigating features for chatgpt. In *International conference on artificial intelligence in education technology* (pp. 152-170). Singapore: Springer Nature Singapore.
6. Wang, R. E., Ribeiro, A. T., Robinson, C. D., Loeb, S., & Demszky, D. (2024). Tutor CoPilot: A human-AI approach for scaling real-time expertise. *arXiv preprint arXiv:2410.03017*.
7. Maadi, M., Akbarzadeh Khorshidi, H., & Aickelin, U. (2021). A review on human–AI interaction in machine learning and insights for medical applications. *International journal of environmental research and public health*, *18*(4), 2121.
8. Dai, C. P., & Ke, F. (2022). Educational applications of artificial intelligence in simulation-based learning: A systematic mapping review. *Computers and Education: Artificial Intelligence*, *3*, 100087.
9. Khuat, T. T., Kedziora, D. J., & Gabrys, B. (2023). The roles and modes of human interactions with automated machine learning systems: A critical review and perspectives. *Foundations and Trends® in Human–Computer Interaction*, *17*(3-4), 195-387.
10. Huang, Z., Sheng, Z., & Chen, S. (2024). Trustworthy human-ai collaboration: Reinforcement learning with human feedback and physics knowledge for safe autonomous driving. *arXiv preprint arXiv:2409.00858*.

11. Glickman, M., & Sharot, T. (2025). How human–AI feedback loops alter human perceptual, emotional and social judgements. *Nature Human Behaviour*, *9*(2), 345-359.
12. Islam, M. S., Das, S., Gottipati, S. K., Duguay, W., Mars, C., Arabneydi, J., ... & Taylor, M. E. (2025). Human-AI collaboration in real-world complex environment with reinforcement learning. *Neural Computing and Applications*, 1-31.
13. Block, S., Nyhuis, D., Gross, M., Harmening, M., & Velimsky, J. A. Classifying Documents with Human-AI-Collaboration: Introducing the Human-AI Collaboration in Classification Utility Framework.