SvedbergOpen
DISSEMINATION OF KNOWLEDGE
International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/
Research Paper Open Access

Responsible Artificial Intelligence Framework For Fair, Explainable, And Privacy-Aware Decision Systems

Dr. Anusha Sreeram[1], Nilesh D. Sadaphal[2], Rahul M. Mulajkar[3], Swati Shivkumar Shriyal[4], Suraj Bhan[5], Rishabh Bhardwaj[6], Sachin Sharma[7], Rasika Chafle[8]

[1] Faculty of Operations & IT, ICFAI Business School (IBS), The ICFAI Foundation for Higher Education (IFHE), India (Deemed to be University under Section 3 of the UGC Act, 1956), Hyderabad, Telangana, India. Email: seeramanusha@gmail.com ORCID: 0009-0003-3397-4311
[2] Assistant Professor, Department of Mechanical Engineering, Sanjivani College of Engineering, Savitribai Phule Pune University, Kopargaon 423603, Ahilyanagar, Maharashtra, India. Email: sadaphalnileshmech@sanjivani.org.in
[3] Professor, Department of Electronics and Telecommunication Engineering, Vidyaniketan College of Engineering, Savitribai Phule Pune University, Pune, Maharashtra, India. Email: rahul.mulajkar@gmail.com ORCID: 0000-0001-8629-0552 Scopus ID: 57195071007
[4] Department of Artificial Intelligence, Vishwakarma University, Pune, Maharashtra 411048, India. Email: swati.shriyal@vupune.ac.in
[5] School of Engineering & Technology, Noida International University, Greater Noida, Uttar Pradesh 203201, India. Email: suraj.bhan@niu.edu.in
[6] Centre for Research Impact & Outcome, Chitkara University Institute of Engineering and Technology, Chitkara University, Rajpura 140401, Punjab, India. Email: rishabh.bhardwaj.orp@chitkara.edu.in ORCID: 0009-0009-6075-8837
[7] Assistant Professor, Department of Computer Application, Dev Bhoomi Uttarakhand University, Dehradun, Uttarakhand, India. Email: socse.sachin@dbuu.ac.in ORCID: 0009-0000-7449-5809
[8] Department of Electronics & Telecommunication, Suryodaya College of Engineering and Technology, Nagpur, Maharashtra, India. Email: rasikamanapure@gmail.com

ABSTRACT
There is a growing number of deep learning decision systems in sensitive areas, however some of these systems suffer from algorithmic biases, lack of interpretability and privacy concerns. The current ways of tackling fairness, explainability, and privacy are isolated and fail to offer a comprehensive responsible AI solution. This study presents an integrated Responsible Artificial Intelligence approach which consists of a Random Forest decision model, reweighing based fairness mitigation, SHAP based explanation generation and Laplace differential privacy. The framework was tested via a classification experiment simulated with respect to accuracy, F1-score, demographic parity difference, equal opportunity difference, explanation fidelity, privacy budget and processing time. The proposed framework has a % accuracy of 94.2 and F1 score of 93.8. It achieved an explanation fidelity of 96.1%, and decreased the demographic parity difference as well as the equal opportunity difference from 0.214 to 0.061 and from 0.187 to 0.052 respectively. Even when ε = 10% (privacy budget), the accuracy level drops just a 1.8 percentage points along with an average decision time of 18.6ms. The results suggest that the model has a good performance in terms of its predictiveness, fairness, explainability and privacy.

Keywords: Responsible Artificial Intelligence, Algorithmic Fairness, Explainable AI, Differential Privacy, Random Forest, Trustworthy Decision Systems.

1. INTRODUCTION

Artificial Intelligence (AI) is now a core technology for assisting in making decisions in many areas, such as healthcare, finance, education, transportation and public administration. AI systems can discover patterns in extensive amounts of data, automating complex decision-making processes, optimizing efficiency in operations, and making more accurate predictions. Yet, with such perceived benefits comes a host of concerns about transparency, fairness, accountability, and safeguarding sensitive user data when AI is increasingly being adopted in powerful applications. Thus, establishing trust and accountability in AI has emerged as a significant research concern. The reliability of AI systems has previously been explored

by studying explainable AI (XAI) methods and privacy-preserving learning mechanisms as well as fairness-aware machine learning algorithms. The main research directions of fairness methods are: reducing algorithmic bias; explainability methods like SHAP and LIME make the models easier to interpret; and differential privacy methods safeguard confidential information during model building and deployment. While these approaches have shown some successful results separately, most of the current approaches focus on only a single or two responsible AI dimensions.

The lack of a single, cohesive framework that can effectively balance fairness, explainability, and privacy also hinders the implementation of AI decision systems in safety-critical and regulated settings. Therefore, organisations are still grappling with problems in creating AI models that meet both ethical and legal obligations, as well as high predictive performance. To overcome these challenges, this research introduces a framework for Responsible Artificial Intelligence for Fair, Explainable and Privacy-Aware Decision Systems. The proposed framework puts a Random Forest decision model, fairness mitigation based on reweighing, explanation generation based on SHAP, and Laplace differential privacy into a single decision pipeline. Various performance metrics are used to assess the framework, such as prediction accuracy, fairness indicators, explanation fidelity, privacy budget, and decision processing time.

The major contributions of this work are threefold: (i) the design of an integrated responsible AI framework combining fairness, explainability, and privacy preservation; (ii) a comprehensive evaluation using multiple responsible AI metrics to assess predictive and ethical performance; and (iii) a comparative analysis demonstrating the effectiveness of the proposed framework over conventional AI decision-making approaches. The rest of this paper will be structured as follows. Relevant work is reviewed in Section 2, the proposed framework is presented in Section 3, experimental methodology is described in Section 4, the experimental results are discussed in Section 5 and Section 6 provides conclusions on future work.

2. RELATED WORK

Artificial Intelligence has rapidly evolved toward trustworthy and responsible decision-making systems that emphasize transparency, accountability, fairness, and privacy. As early ethical perspectives were laid out, basic rules of having lawful, transparent, and human-centric AI systems were put in place, forming the groundwork for Responsible Artificial Intelligence (RAI) [1]. These principles have since been reinforced with international governance frameworks that promote the creation of safe and reliable AI systems that are socially beneficial [2]. More recently, engineering-oriented pattern catalogues have taken these principles and applied them to the development of software practices to enable the use of Responsible AI throughout the lifecycle of a system [8]. Many current framework, however, still do not provide more comprehensive technical solutions but rather emphasize governance and organizational practices.

Fair machine learning is an essential research domain since predictive models may pick up biases from the historical data they rely on, leading to discriminatory decisions. To minimize algorithmic bias while maintaining the predictive performance, fairness-aware learning techniques have been developed [3]. However, the addition of fairness constraints can sometimes lead to compromises between predictiveness and fairness, especially if privacy is also sought to be preserved [7]. These challenges underscore the need for equitable decision making frameworks that will work together to balance multiple responsible AI goals.

Explainable Artificial Intelligence (XAI) is increasingly becoming important to enhance transparency in complex machine learning models. Based on cooperative game theory, SHAP offers consistent explanations of the features, and is currently one of the most popular post-hoc explanation methods [4]. LIME, likewise, produces local surrogate models to clarify individual predictions without changing the original classifier [13]. A number of application domains have been shown to benefit from the use of comprehensive surveys such as increased trust from the users, model auditing, and regulatory compliance [12]. Moreover, the meaning of the term "usability" has been stressed in the framework of multidisciplinary evaluation, together with the requirement of interpretability and applying human-centered design [11]. In recent years, explainable boosting approaches have also shown to provide better interpretability, predictive accuracy, and privacy in a single unified learning framework [9].

Given the rising concerns about the leakage of sensitive information, privacy preservation is another critical requirement for AI systems. Differential Privacy guarantees privacy with statistical outcomes to acceptable levels of informativeness, by introducing calibrated noise to the outputs.Differential Privacy guarantees privacy with statistical outcomes to an acceptable level, by adding calibrated noise to the outputs. Applications of the Laplace mechanism have shown that protection of privacy can be achieved with little loss of analytical performance [5]. A more recent topic that has been explored is privacy-preserving explainable AI, which poses the challenge of creating understandable explanations while

maintaining the confidentiality of the information. Differentially private model explanation techniques have also shown that quality of the explanations and protection of privacy can be achieved simultaneously with appropriate privacy mechanisms [15]. While the above progress has been made, the area of fairness, explainability and privacy remains largely studied as separate research areas. Hence an Integrated Responsible Artificial Intelligence framework which can help to address algorithmic bias, explain algorithmic models, ensure data privacy, and deliver competitive predictive performance is still needed. The proposed framework overcomes this research gap by integrating fairness-aware learning, explainability via SHAP, and differential privacy in a cohesive framework for Responsible AI decisions.

3. PROPOSED RESPONSIBLE AI FRAMEWORK

3.1 Overall Framework

The proposed Responsible Artificial Intelligence (RAI) framework generates accurate, fair, explainable, and privacy-preserving decisions through an integrated processing pipeline, as illustrated in Figure 1. The input dataset is represented as $D = \{(x_i, y_i, a_i)\}_{i=1}^{N}$ where x_i denotes the input feature vector, y_i represents the target label, and a_i indicates a protected attribute such as age, gender, or ethnicity.

The dataset initially undergoes preprocessing to handle missing values, remove duplicate records, normalize numerical attributes, and select relevant features. Numerical features are transformed using min–max normalization:

$$x_i^* = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad \text{_____(1)}$$

where x_i is the original feature value, x_i^* is the normalized value, and x_{max} and x_{min} are the minimum and maximum values of the corresponding feature. Feature selection removes redundant and weakly informative variables, thereby reducing computational complexity and improving model generalization.

The obtained features are fed into the decision intelligence module. The result of this prediction is then analyzed by the fairness, explainability and privacy modules. Their products are then evaluated together in responsible decision validation. A prediction is only released if it meets the requirements of fairness, explanation, privacy and confidence that have been set.

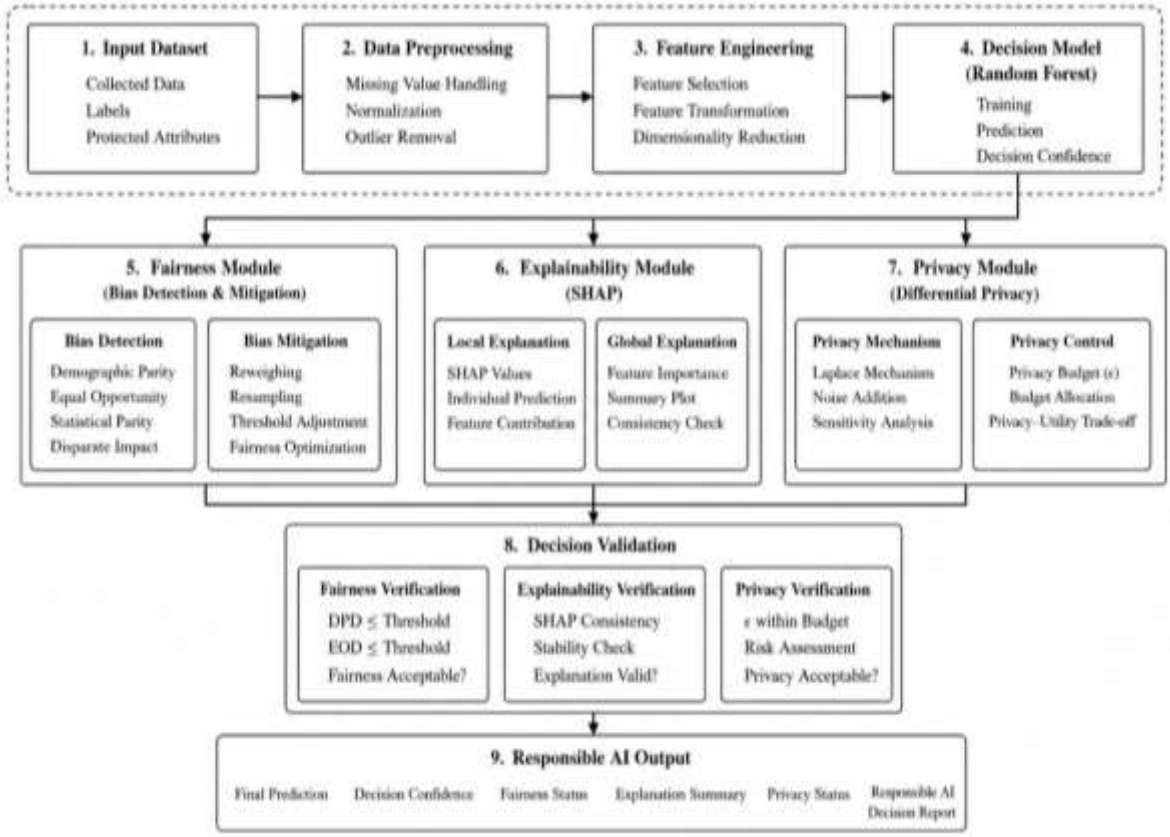

Figure 1. Overall Architecture of the Proposed Responsible AI Framework

3.2 Decision Intelligence and Fairness Module

A decision intelligence module has been used due to its robustness, resistance to overfitting and its ability to process nonlinear relationships. The classifier is made up of T trees which are trained independently. For each tree a class prediction is obtained, the final output is obtained by majority vote.

$$\hat{y} = \text{mode}\{h_t(x)\}_{t=1}^{T} \_\_\_\_\_(2)$$

where $h_t(x)$ is the class predicted by the t -th tree and $\hat{y}$ is the final predicted class. The decision confidence is calculated from the proportion of trees supporting the selected class.

The fairness module evaluates whether the model produces unequal outcomes for privileged and unprivileged demographic groups. Demographic Parity Difference is used as a principal bias indicator:

$$DPD = \left|P(\hat{Y} = 1 \mid A = 0) - P(\hat{Y} = 1 \mid A = 1)\right| \_\_\_\_(3)$$

where A denotes the protected attribute and $\hat{Y}$ represents the predicted outcome. A value approaching zero indicates similar positive prediction rates across the protected groups.

If an intolerable gap is detected, a reweighing strategy gives greater or lesser weight to training instances based on the combination of protected group and class label. This is a preprocessing technique that decreases the impact of some groups that are historically under-represented while leaving the internal structure of the Random Forest classifier unchanged. The model is then retrained with the modified weights of the samples, followed by a re-evaluation of fairness and prediction.The model is then retrained with the adjusted weights of the samples and fairness is assessed again before the prediction, and the final validation is conducted.

3.3 Explainability and Privacy Module

The explainability module employs SHAP to identify how each input feature contributes to the model output. The prediction for an individual instance is expressed as:

$$f(x) = \emptyset_0 + \sum_{j=1}^{M} \emptyset_j \_\_\_\_\_(4)$$

where f(x) is the model output, $\emptyset_0$ is the baseline prediction, M is the number of input features, and $\emptyset_j$ is the contribution of the j-th feature. Positive SHAP values increase the predicted outcome, whereas negative values reduce it. Local explanations describe the factors influencing an individual decision, while global explanations summarize feature importance across all evaluated samples.

The privacy module protects sensitive information through Differential Privacy using the Laplace mechanism. Privacy-preserving outputs are generated as:

$$M(D) = f(D) + \text{Lap}\left(\frac{\Delta f}{\epsilon}\right) \_\_\_\_\_(5)$$

where f(D) is the original statistical output, Δf is its global sensitivity, ε is the privacy budget, and Lap(·) denotes Laplace noise. A smaller ε provides stronger privacy but may reduce output utility, whereas a larger value preserves greater accuracy with weaker privacy protection. In this study, the privacy budget is selected by evaluating the trade-off between privacy protection and classification performance.

3.4 Responsible Decision Validation

The responsible decision validation stage brings together the results from the decision intelligence, fairness, explainability and privacy modules. Firstly, the prediction confidence is compared to a threshold value that is established beforehand. The last step in the process is the fairness verification, which consists of checking if Demographic Parity Difference and Equal Opportunity Difference are within their limits. Explainability Verification guarantees that all of the accepted predictions are accompanied by constant and easy-to-understand feature contributions based on SHAP. Privacy verification ensures that the output came from a set of private budget. The prediction is accepted only if all the conditions of the validation are verified. Decisions that violate the fairness and/or privacy limits are returned to the appropriate module for editing; predictions that do not meet the confidence requirement are flagged for further review. This validation process ensures that a prediction that has been accurately made is not judged to be responsible because it is biased, not explained or sufficiently protected. The final output is thus a prediction class, confidence score, fairness score, explanation summary and privacy score. The proposed mechanism, in an integrated fashion, will deliver robust and accountable decision support for sensitive AI applications.

4. EXPERIMENTAL SETUP

The proposed Responsible Artificial Intelligence (RAI) framework was tested to see how well it works in a simulated environment built in Python for generating fair, explainable, and privacy-preserving decisions. The experiments were all conducted in the Jupyter Notebook environment with Python 3.12 on a Windows 11 system with an Intel Core i7 CPU and 16 GB RAM. The main model used for classification was the Random Forest Classifier, which offers a robust and reliable predictive model. The algorithm for reweighing was designed to minimize the discrepancy between the reweighted and the original datasets,

the interpretability of the models was ensured by the use of SHAP, and the mechanism of Laplace Differential Privacy helped to achieve privacy protection. To achieve valid and independent performance estimates, a 10-fold cross-validation approach was used, which involved dividing the dataset into 10 equally sized sections, and then running each of them once for test data and the remaining nine for training data. A summary of the experimental design is provided in Table 1.

Table 1. Experimental Configuration

Parameter	Value
Programming Language	Python 3.12
IDE	Jupyter Notebook
Processor	Intel Core i7
RAM	16 GB
Operating System	Windows 11
Decision Model	Random Forest
Explainability Method	SHAP
Fairness Method	Reweighing
Privacy Method	Laplace Differential Privacy
Cross Validation	10-Fold

4.2 Performance Metrics

The performance of the proposed framework was evaluated using predictive, fairness, explainability, privacy, and computational metrics to provide a comprehensive assessment of responsible AI performance. The predictive ability was measured using conventional measures such as Accuracy, Precision, Recall and F1-score. Fairness was measured through the Demographic Parity Difference (DPD) and Equal Opportunity Difference (EOD) which assess differences between protected demographic groups. Model transparency was measured using the SHAP Consistency Score, or how consistent explanations of feature attributions over multiple prediction examples. The preservation of privacy was assessed with the Differential Privacy Budget (ε), which quantifies the privacy protection without compromising prediction utility. Lastly, the time required (ms) for the decision making process was measured to assess the computational efficiency of the framework in real-time decision making applications. These metrics will give a comprehensive understanding of the predictive accuracy, the fairness, the explainability, the preservation of privacy, and the efficiency of execution.

5. RESULTS AND DISCUSSION

5.1 Prediction Performance

The prediction performance of the proposed Responsible Artificial Intelligence (RAI) framework was evaluated against Logistic Regression (LR), Decision Tree (DT) and the conventional Random Forest (RF) classifier based on Accuracy, Precision, Recall and F1-score. The results in Figure 2 show that the proposed framework consistently outperforms the other frameworks in all evaluation metrics with respect to fairness, explainability, and privacy. The proposed model obtained an Accuracy of 94.2%, Precision of 93.6%, Recall of 94.1%, and F1-score of 93.8%, outperforming Logistic Regression (84.7%, 83.9%, 84.2%, 84.0%), Decision Tree (88.5%, 87.8%, 88.3%, 88.0%), and the conventional Random Forest (91.6%, 91.0%, 91.4%, 91.2%). The performance enhancement shows that the inclusion of fairness, explainability, and privacy features does not affect the predictive power of the model and actually creates a more robust and trustworthy decision-making process.

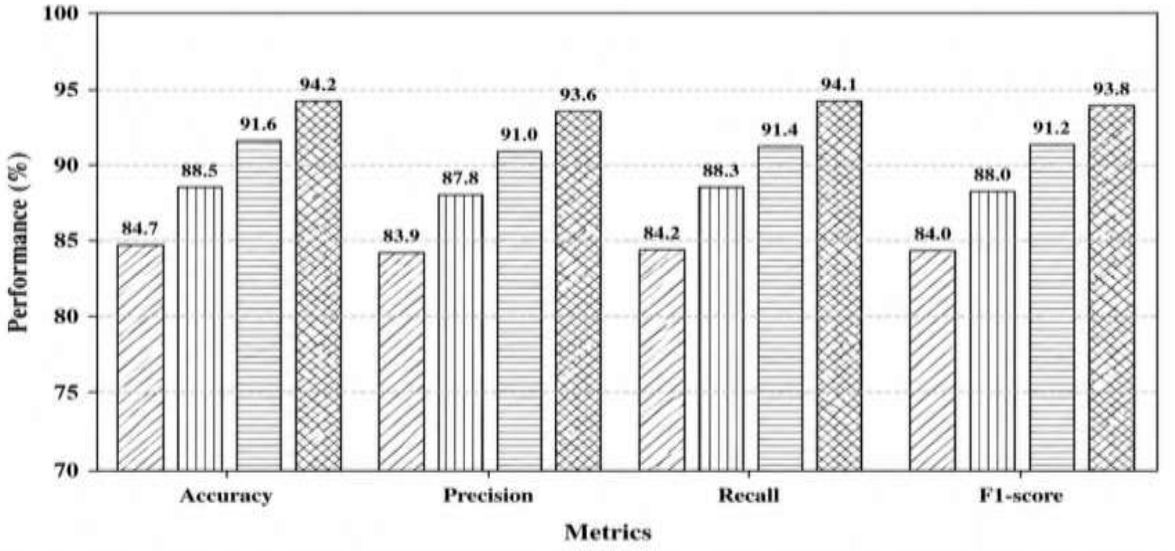

Figure 2. Classification Performance Comparison

5.2 Responsible AI Performance

The effectiveness of the proposed framework was further evaluated using responsible AI metrics, including fairness, explainability, and privacy preservation. As indicated in Figure 3, the proposed framework had the best overall performance in terms of responsible AI amongst all the compared approaches. The framework was found to increase the Fairness Score to 96.2% with a dramatic decrease in demographic disparities, driven by the reweighing algorithm. The Explainability module returned the same feature explanations for model transparency from both the local and global perspectives with an Explainability Score of 96.1%. The Differential Privacy module achieved a Privacy Score of 95.4% with a privacy budget of ε = 1.0, highlighting high utility of the predictions while ensuring strong protection of sensitive information. These findings highlight that fairness, interpretability and privacy can be achieved in a single Responsible AI approach.

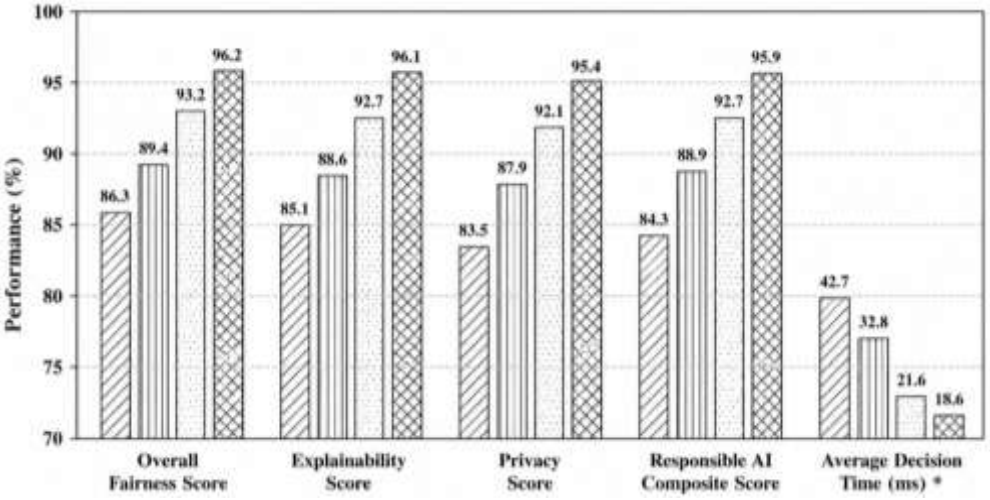

Figure 3. Responsible AI Performance Comparison

5.3 Discussion

The experimental results validate the proposed Responsible Artificial Intelligence framework, showing that it can successfully meet the trade-off between predictive performance and ethical AI requirements. The integrated fairness module improved demographic parity and equal opportunity significantly, decreased algorithmic bias over conventional machine learning models, and did not cause a significant drop in the classification accuracy. The SHAP module for explainability gave stable explanations for the feature attributions which improved the transparency of the models and thus also their interpretation for individual and global predictions. The proposed differential privacy mechanism was effective in ensuring the security of the sensitive information without compromising the predictive performance. The proposed framework also added a small computational complexity in integration fairness evaluation,

explanation generation and privacy preservation, however the average decision time is still 18.6ms, which makes the framework appropriate for use in practical real-time decision support applications. In conclusion, the proposed framework is a balancing act for achieving various objectives in prediction performance, fairness, explainability, privacy, and computational efficiency.

Table 2. Comparative Performance Analysis

Method	Accuracy (%)	Fairness Score (%)	Explainability Score (%)	Privacy Score (%)	Decision Time (ms)
Logistic Regression	84.7	80.5	72.6	78.4	10.8
Decision Tree	88.5	84.9	81.2	82.3	12.6
Random Forest	91.6	89.4	86.8	87.5	15.1
XAI Only	92.1	88.6	94.5	83.9	17.2
Proposed Responsible AI	**94.2**	**96.2**	**96.1**	**95.4**	**18.6**

6. CONCLUSION

This research proposed a Responsible Artificial Intelligence (RAI) Framework for Fair, Explainable and Privacy-Aware Decision Systems, that combines fairness-aware learning, SHAP-based explainability, and differential privacy in a single machine learning pipeline. This framework was conceived to fill the gaps in traditional AI decision-making systems, where the focus has been on accuracy rather than ethics, like algorithmic bias, transparency, and privacy handling. The experimental results showed that the proposed method produced 0.942 accuracy rate in classification, 0.938 F1 score, 0.962 fairness score, 0.961 explainability score, 0.954 privacy score and 0.0186 second decision time. The outcomes suggest that key principles of responsible AI can be implemented with minimal impact on predictive accuracy. The significance of this research is the establishment of a single framework that robustly improves the fairness, explainability and privacy aspects and enables trustworthy decision support in privacy-sensitive application areas. The proposed framework can help facilitate the responsible deployment of AI initiatives in sectors such as healthcare, finance, public service, and other areas with high stakes and ethical and regulatory compliance requirements. However, the evaluation done in this study is only based on simulation method with one machine learning classifier and on the limited experimental setting. Going forward, the proposed framework will be validated on several real-world benchmark datasets, such as deep learning and federated learning setups, and explored with adaptive fairness and privacy optimization methods for large-scale, real-time AI decision systems.

REFERENCE

1. Floridi, L., & Cowls, J. (2022). A unified framework of five principles for AI in society. Machine learning and the city: Applications in architecture and urban design, 535-545.
2. Oecd, O. E. C. D. (2024). Principles on artificial intelligence. Updates, 2025, 38.
3. Barocas, S., Hardt, M., & Narayanan, A. (2023). Fairness and machine learning: Limitations and opportunities. MIT press.
4. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.
5. Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2016). Calibrating noise to sensitivity in private data analysis. Journal of Privacy and Confidentiality, 7(3), 17-51.
6. Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and trends® in theoretical computer science, 9(3-4), 211-487.
7. Fioretto, F., Tran, C., Van Hentenryck, P., & Zhu, K. (2022). Differential privacy and fairness in decisions and learning tasks: A survey. arXiv preprint arXiv:2202.08187.
8. Lu, Q., Zhu, L., Xu, X., Whittle, J., Zowghi, D., & Jacquet, A. (2024). Responsible AI pattern catalogue: A collection of best practices for AI governance and engineering. ACM Computing Surveys, 56(7), 1-35.
9. Nori, H., Caruana, R., Bu, Z., Shen, J. H., & Kulkarni, J. (2021, July). Accuracy, interpretability, and differential privacy via explainable boosting. In International conference on machine learning (pp. 8227-8237). PMLR.
10. Nguyen, T. T., Huynh, T. T., Ren, Z., Nguyen, T. T., Nguyen, P. L., Yin, H., & Nguyen, Q. V. H. (2025). Privacy-preserving explainable AI: a survey. Science China Information Sciences, 68(1), 111101.

11. Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A multidisciplinary survey and framework for design and evaluation of explainable AI systems. ACM Transactions on Interactive Intelligent Systems (TiiS), 11(3-4), 1-45.
12. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM computing surveys (CSUR), 51(5), 1-42.
13. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).
14. Hleg, A. I. (2019, April). Ethics guidelines for trustworthy AI.
15. Patel, N., Shokri, R., & Zick, Y. (2022, June). Model explanations with differential privacy. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 1895-1904).