IS 596
Designing Intelligent Systems

AI-Powered Hiring Decision Support: Enhancing Bias Detection and Decision-Making through Explainable AI

1 Introduction

Artificial intelligence (AI) is transforming hiring processes by offering faster and more scalable candidate evaluations. However, AI systems trained on biased historical data risk perpetuating existing inequalities. This paper introduces an AI-powered Hiring Decision Support System (HDSS) designed to enhance fairness and transparency in hiring. The system combines predictive modeling, bias detection, and natural language explanations generated by large language models (LLMs) to help recruiters make equitable and well-informed decisions.

2 Research Questions

This study explores the intersection of bias detection and explainable AI in hiring contexts through the following research questions:

1. How can predictive hiring models and LLMs be integrated to detect and communicate bias in candidate evaluations?
2. In what ways can explanation design improve recruiter understanding and trust in AI-generated hiring recommendations?
3. Can combining bias detection with explainability enhance decision-making and reduce unfair outcomes in hiring?

3 Literature Review

3.1 AI Bias in Hiring

Several studies have documented bias in AI hiring systems. Raghavan et al. (2020) revealed that without intervention, commercial hiring tools often replicate historical biases. Amazon's AI tool, for instance, was discontinued after it favored male candidates, due to training on skewed historical data (Dastin, 2018).

3.2 Explainable AI (XAI) in Hiring

XAI seeks to make AI systems interpretable and trustworthy. SHAP (Lundberg & Lee, 2017) and LIME are widely adopted feature attribution methods that quantify the influence of each input variable on model predictions. However, while technically rigorous, their numeric outputs are often inaccessible to HR professionals. Beretta et al. (2024) emphasize that textual explanations are more effective in hiring settings, enabling recruiters to reason about AI decisions.

3.3 Cognitive Biases in Decision-Making

Human hiring decisions are subject to cognitive biases such as the halo effect, affinity bias, and confirmation bias (Tversky & Kahneman, 1974). These biases can be reinforced by opaque AI models. Hofeditz et al. (2022) argue that structured explanations and fairness indicators serve as cognitive debiasing tools, prompting recruiters to question biased patterns and consider underrepresented candidates more equitably.

3.4 Role of LLMs in Explanation Generation

Large language models like GPT-4 and Mistral-7B are increasingly used to interpret model outputs. Zytek et al. (2024) propose that LLMs not only improve interpretability but can also contextualize model decisions in plain language. Their flexibility allows explanation style to be tailored for different users—whether technical analysts or business decision-makers.

3.5 Mental Models and Human-AI Interaction

Norman (1983) defines mental models as the cognitive structures people use to understand complex systems. In AI hiring, if recruiters cannot form accurate mental models of how predictions are made, they may over-rely on or completely distrust the system. Gregor and Benbasat (1999) argue that explanations should support these mental models by providing clear, relevant, and user-aligned information.

Overall, this literature underscores the importance of embedding interpretability and fairness directly into the design of hiring algorithms. Our system builds on these insights by combining predictive accuracy with cognitive-aligned explanations and fairness auditing.

4 Study Rationale

Despite the growing use of AI in hiring, many systems lack transparency and bias mitigation capabilities. This study addresses that gap by developing a system that detects bias and explains its decisions in ways aligned with how recruiters think. The goal is not to replace human judgment but to support it with interpretable and fair algorithmic insights.

5 Study Approach

To address the dual challenges of fairness and interpretability in AI-assisted hiring, we adopted a modular and iterative design approach to build the Hiring Decision Support System (HDSS). This involved synthesizing a pipeline that mimics real-world hiring evaluations while remaining controllable for experimental purposes.

We began by creating a synthetic dataset of 50 candidate profiles, carefully designed to reflect realistic variation in hiring qualifications, such as GPA, years of professional experience, education level, certifications, and project involvement. Gender was included as a demographic attribute to test bias detection. Synthetic data was chosen due to the unavailability of public hiring datasets and allowed us to simulate a diverse applicant pool while avoiding privacy concerns. To maintain data quality, we ensured balanced distribution and representation across demographic groups.

The system's development followed a three-tier pipeline architecture:

5.1 Predictive Modeling Layer: A Random Forest Regressor was selected after comparative testing with linear regression and support vector machines. Random Forest provided a strong balance between performance and interpretability, particularly when used in conjunction with SHAP for feature attribution.

5.2 Bias Detection Layer: This component performs post-hoc auditing on model outputs. It analyzes predicted hire scores for group-wise discrepancies—initially across gender. This layer is modular, enabling future plug-in of fairness metrics such as demographic parity or equalized odds.

5.3 Explanation Layer: To bridge the technical-human interpretability gap, we developed a local explanation generator that integrates SHAP-derived feature contributions with a Mistral-7B language model running via Ollama. This local deployment was selected to ensure data privacy and minimize latency.

All components were written in Python and developed to run in a reproducible environment. The system was tested under various simulated edge cases, such as outlier candidate profiles, to evaluate robustness. Additionally, we designed the pipeline to be extensible, so additional features (e.g., age or race) and explanation modes (e.g., counterfactuals) could be easily incorporated.

This approach allowed us to iteratively improve both the predictive accuracy and the quality of explanations while ensuring the system remained auditable, transparent, and adaptable to future needs.

6 Technical Explanation of the System

The HDSS system consists of three primary technical components: the AI Hiring Model, the Bias Detection Module, and the Explanation Generator. Together, these components form a continuous workflow from raw input to actionable recruiter insight.

6.1 AI Hiring Model

Implemented using scikit-learn’s Random Forest Regressor, this model accepts structured inputs such as:

- GPA (float)
- Years of professional experience (int)
- Education level (ordinal categorical)
- Number of certifications (int)
- Number of projects (int)

The target variable is a synthetic "hireability score" derived from assumed best practices in hiring decisions. We selected Random Forest due to its:

- High performance on tabular data

- Ability to handle both categorical and continuous variables
- Natural compatibility with tree-based explainability frameworks (i.e., TreeSHAP)

Hyperparameters such as the number of trees, max depth, and minimum samples per leaf were tuned using grid search and cross-validation on an 80/20 train-test split. The resulting model achieved an R^2 score of 0.82, suggesting strong internal validity within the pilot dataset.

6.2 Bias Detection Module

This component performs group fairness audits by evaluating the predicted scores across demographic slices. It begins by computing group-wise descriptive statistics (mean, variance) for hire scores and then applies thresholds or statistical tests (e.g., two-sample t-tests) to flag significant discrepancies.

In our current implementation, we focused on gender-based analysis across three categories: male, female, and non-binary. The system is designed with flexibility in mind: demographic attributes and thresholds can be redefined by users, making the module adaptable to organizational or legal requirements (e.g., EEOC compliance).

Rather than enforcing interventions, this module presents informational diagnostics to recruiters, enhancing transparency without removing agency.

6.3 Explanation Generator

The interpretability engine combines SHAP values and natural language generation. After the predictive model produces a score, SHAP computes a local explanation indicating how much each feature contributed to that score.

We then constructed structured prompt templates that include this information in a format suitable for an LLM. These prompts are passed to a locally hosted Mistral-7B model running via Ollama, which transforms the SHAP output into a recruiter-friendly narrative.

Example: “This candidate’s high score is primarily driven by 11 years of experience and four certifications. Their GPA and master’s degree had a moderate impact on the final score.”

The LLM is sandboxed and runs on-device, preserving data privacy and enabling offline deployment—essential for environments with sensitive candidate data.

In future iterations, we plan to fine-tune the LLM on HR-specific corpora to enhance tone, reduce verbosity, and improve alignment with recruiter language styles.

7 Use Cases and Scenarios of System Testing

To test the robustness and relevance of the Hiring Decision Support System (HDSS), we designed and executed a set of simulated use cases that mirror real-world hiring scenarios. Each use case involved diverse candidate profiles, representing varied levels of qualifications and demographic characteristics. The system was tested for predictive behavior, bias detection sensitivity, and clarity of generated explanations.

7.1 Use Case 1: Fair Candidate Evaluation

Use Case: A recruiter wants to rank 50 applicants for a role based on skills and qualifications without introducing unconscious bias.

Your System's Role: Automatically generates hireability scores and flags potential gender-based bias in the results.

7.2 Use Case 2: Transparent Decision Justification

Use Case: A hiring manager needs to justify why a shortlisted candidate received a higher score than others during a hiring panel discussion.

Your System's Role: Provides SHAP-based explanations and a natural language summary from the LLM to support the hiring decision.

7.3 Use Case 3: Bias Monitoring Over Time

Use Case: An HR department wants to ensure fairness across multiple recruitment cycles and check for long-term trends in bias.

Your System's Role: Logs and analyzes predicted scores across demographic groups over time for bias trends.

7.4 Use Case 4: Candidate Comparison

Use Case: Two equally qualified candidates are being considered. The recruiter wants AI input to decide based on skill and experience.

Your System's Role: Offers a side-by-side analysis and a recommendation with reasoning from the LLM.

8 Evaluation and Results

We evaluated the HDSS on several dimensions: model performance, explanation quality, bias detection sensitivity, and system usability.

8.1 Model Performance

The Random Forest Regressor achieved an **R^2 score of 0.82** on the holdout test set of synthetic candidates, suggesting strong predictive accuracy within the defined feature space. We selected Random Forest after comparing performance with linear regression and decision tree baselines, which underperformed in capturing non-linear patterns in the data. The model exhibited low variance and high consistency across multiple resampling runs.

8.2 Explanation Quality

We generated explanations for five randomly selected candidates and performed a qualitative analysis. Each explanation was compared against SHAP values and manually verified for alignment. Four out of five explanations correctly represented the top contributing features and their relative influence. One explanation emphasized experience more than the actual SHAP values suggested, indicating a mild discrepancy in LLM interpretation. This shows the need for better prompt calibration or tighter integration of SHAP metrics into the LLM prompt structure.

Explanations were evaluated for:

- **Clarity**: understandable to a non-technical recruiter
- **Faithfulness**: alignment with SHAP values
- **Length and Tone**: professional, concise, and informative

Recruiters in simulated review sessions reported improved understanding of why candidates received specific scores, and appreciated the natural language format.

8.3 Bias Detection

We tested bias detection by:

1. Running profiles of similar candidates with only the gender attribute changed.
2. Manually manipulating datasets to introduce score gaps for specific groups.

In both tests, the system flagged average score disparities above 5% as potential bias. While the current setup only supports gender, the architecture is modular and supports expansion to race, age, and other demographics.

8.4 System Usability

The end-to-end processing time (from CSV input to explanation output) was approximately 4–6 seconds per candidate on a local machine. The system was lightweight enough to be deployed on a laptop, making it suitable for small-scale HR departments. Outputs were formatted clearly, with JSON and plain-text export options, supporting integration into future UI designs.

In sum, our evaluation shows that the HDSS performs well as a prototype, offering strong predictive accuracy, actionable explanations, and effective bias detection—all within a reproducible and transparent framework.

9 Discussion

The results from testing and evaluation suggest that the HDSS is not only a technically sound system but also a conceptually valuable tool for ethical hiring. By combining algorithmic predictions with fairness monitoring and interpretable explanations, it addresses multiple pain points in AI-assisted decision-making.

9.1 Human-Centered Design and Trust

A key strength of the HDSS lies in its support for **human-centered AI**. The system does not automate hiring but assists recruiters by providing transparent and contextualized predictions.

This approach mitigates both blind trust (i.e., overreliance on AI) and distrust (i.e., rejection of AI tools), which are common when users don't understand algorithmic behavior.

9.2 Ethical Implications and Bias Mitigation

Bias detection is particularly crucial in high-stakes domains like hiring. Our implementation offers algorithmic auditing capabilities that surface disparities across demographic groups. Rather than masking these issues, the system foregrounds them, encouraging recruiters to consider fairness implications.

Still, the current implementation has limitations. Using a synthetic dataset, while necessary for privacy and control, may not capture the full variability and complexity of real-world data. The bias module currently only addresses gender, and the explanation generator, while effective, may not perfectly reflect the internal logic of the model unless SHAP-LLM integration is improved.

9.3 Future Directions

While the HDSS prototype successfully demonstrates the feasibility of integrating bias detection and explainable AI into hiring workflows, there are several critical opportunities for future development that would enhance its scalability, generalizability, and real-world impact.

Integration with Real-World Datasets

To fully validate the system's robustness, testing on real-world hiring data is essential. Collaborations with HR departments or anonymized corporate datasets would allow us to evaluate the model's performance in environments with greater feature variability, missing values, and noise. This step would also help refine the bias detection thresholds and improve generalizability.

Expanded Demographic Fairness Auditing

Currently, the bias detection module is limited to analyzing gender disparities. In future iterations, the system will incorporate fairness assessments across other sensitive attributes such as race, age, and disability status. Additionally, intersectional bias detection (e.g., examining bias against women of color specifically) will be introduced to provide a more nuanced understanding of systemic disparities. Integration of fairness metrics like equalized odds, disparate impact ratio, and demographic parity can further quantify the degree of bias.

Enhanced Explanation Fidelity and Personalization

Although our explanation module delivers coherent narratives, it occasionally lacks precise alignment with SHAP value magnitudes. Future work will involve tighter coupling between SHAP outputs and the LLM prompt structure—potentially using prompt engineering techniques or fine-tuning the LLM on explanation-specific training data. Moreover, explanation tone and format could be personalized based on user roles. For example, HR professionals might prefer high-level summaries, while compliance officers may require more granular, metric-based breakdowns.

Scalability and Real-Time Performance Optimization

To support real-time or high-volume hiring environments, the system's latency must be reduced. We plan to explore lighter-weight LLM architectures, asynchronous explanation pipelines, or batching mechanisms for SHAP and generation processes. In enterprise contexts, integration with cloud infrastructure (e.g., AWS SageMaker or Azure ML) could further improve scalability and deployment reliability.

User Interface and Experience Design

Currently, HDSS operates through scripts and console outputs. Future development will focus on building an interactive user interface (UI) with visual dashboards, candidate cards, explanation modals, and bias alerts. Usability testing with recruiters will guide UI design decisions to ensure clarity, interpretability, and minimal cognitive load during hiring evaluations.

Human-in-the-Loop Feedback and Continuous Learning

We also plan to integrate a feedback mechanism where recruiters can rate the usefulness of each explanation or flag instances where the model's reasoning feels inconsistent. This feedback could be used to iteratively fine-tune the LLM explanations or retrain the scoring model, enabling the system to evolve based on real-world usage and decision outcomes.

In summary, these future directions aim to elevate HDSS from a research prototype to a deployable, ethical AI tool for hiring. They reflect our commitment to building AI systems that are not only technically sound but also human-centered, legally compliant, and aligned with organizational values.

10. Conclusion

This project presents a prototype Hiring Decision Support System that integrates predictive modeling, bias detection, and explainable AI to support fairer and more transparent hiring practices. By combining a Random Forest model with SHAP-based feature attribution and natural language explanations from a local LLM, the system not only predicts hireability scores but also makes the rationale behind each decision accessible to human recruiters.

The HDSS highlights potential demographic disparities and delivers recruiter-friendly justifications, helping to build trust and accountability in algorithmic recommendations. While currently limited by synthetic data and scope of demographic analysis, its modular architecture allows for easy extension to real-world data, expanded fairness metrics, and more refined explanations.

Overall, this work contributes to the growing field of responsible AI in hiring by showing how interpretability and fairness can be embedded directly into system design—empowering human decision-makers without removing them from the loop.

References

Beretta, A., Ercoli, G., Ferraro, A., Guidotti, R., Iommi, A., Mastropietro, A., Monreale, A., Rotelli, D., & Ruggieri, S. (n.d.). *Requirements of eXplainable AI in Algorithmic Hiring*. CEUR Workshop Proceedings. https://ceur-ws.org/Vol-3744/paper8.pdf

Bonde, S., Girish, S., & Bhatambrekar, V. (2025). *AI-powered hiring decision support* [Computer software]. GitHub. https://github.com/dataprojectdis/AIPowered_HiringDecisionSupport

Dastin, J. (2018, October 11). *Amazon scraps secret AI recruiting tool that showed bias against women*. Reuters. https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/

Gregor, S., & Benbasat, I. (1999). Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS Quarterly*, 23(4), 497–530. https://doi.org/10.2307/249487

Hofeditz, L., Clausen, S., Rieß, A., Mirbabaie, M., & Stieglitz, S. (2022). Applying XAI to an AI-based system for candidate management to mitigate bias and discrimination in hiring. *Electronic Markets, 32*(4), 2207–2233. https://doi.org/10.1007/s12525-022-00600-9

Lipton, Z. C. (2018). The mythos of model interpretability. *Queue*, 16(3), 31–57. https://doi.org/10.1145/3236386.3241340

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *arXiv preprint*. https://arxiv.org/abs/1705.07874

Norman, D. A. (1983). Some observations on mental models. In D. Gentner & A. L. Stevens (Eds.), *Mental Models*. Psychology Press. https://doi.org/10.4324/9781315802725

Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic hiring. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 469–481. https://doi.org/10.1145/3351095.3372828

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Zytek, A., Pidò, S., & Veeramachaneni, K. (2024). *LLMs for XAI: Future directions for explaining explanations*. arXiv. https://arxiv.org/abs/2405.06064