



International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Research Paper

Open Access

Explainable Semantic Resume Screening With Fairness-Aware Candidate Ranking

Dr. P. Harini¹, Dr. C. Hari Kishan², Dr. D. Rajendra Prasad³, Dr. K. Jagadeesh Babu⁴, Veera Raghava Swamy Nalluri⁵, Dr. S. Naveen Kumar Polisetty⁶

¹Professor, Department of CSE St. Anns College of Engineering and Technology Chirala, A.P., India. hariniphd@yahoo.com

²Professor, Department of CSE-AIML St. Anns College of Engineering and Technology Chirala, A.P., India. drchkishan@gmail.com

³Professor, Department of ECE St. Anns College of Engineering and Technology Chirala, A.P., India. rp.devathoti@gmail.com

⁴Professor, Department of ECE St. Anns College of Engineering and Technology Chirala, A.P., India. Jagan_ec@yahoo.com

⁵Associate Professor, Department of ECE St. Anns College of Engineering and Technology Chirala, A.P., India. raghavaswamynv@gmail.com

⁶Associate Professor, Department of CSE-AIML St. Anns College of Engineering and Technology Chirala, A.P., India. drnaveenpolisetty@gmail.com

Abstract

Large applicant pools make manual resume screening slow, inconsistent, and difficult to audit. Conventional Applicant Tracking Systems reduce effort through keyword filtering, but exact-term matching can miss semantically equivalent skills and offers limited justification for ranking decisions. This paper presents an explainable semantic screening framework that combines Sentence-BERT-style contextual similarity, structured qualification features, a validation-weighted ensemble of Logistic Regression, Random Forest, and Gradient Boosted Trees, and SHAP-based local explanations. Personally identifying information is separated from model features and retained only for authorized compliance workflows. Candidate ranking is driven by suitability probability, mandatory-skill coverage, and transparent feature contributions rather than raw keyword counts. A controlled synthetic benchmark of 1,760 resume-job pairs across eight role profiles is used to verify the analytical pipeline without presenting unverified real-world hiring outcomes. The proposed ensemble reaches 87.95% accuracy, 90.17% F1-score, and 94.90% ROC-AUC, improving over a keyword/skill baseline. Post-hoc group auditing shows a 1.51 percentage-point equal-opportunity gap in the controlled benchmark. The framework is intended as recruiter decision support, not an autonomous hiring authority, and emphasizes traceability, human override, and repeatable fairness monitoring.

Keywords: Resume Screening, Candidate Ranking, Explainable AI, Sentence Embeddings, Machine Learning, SHAP, Fairness Audit, Human Resources Analytics.

1. Introduction

Digital recruitment has transformed the first stage of hiring from a document-review task into a large-scale information retrieval problem. Recruiters may receive hundreds or thousands of resumes for a single position, while the information that matters—skills, project evidence, work history, education, certifications, and role progression—is expressed in highly variable language. A candidate may write “container orchestration” while the job description specifies “Kubernetes,” or describe “predictive modeling” instead of the exact tool name expected by a keyword filter. Systems that rely primarily on literal term occurrence therefore risk false rejection even when a resume and job description are semantically aligned.

Modern language representations can reduce this lexical mismatch. Contextual encoders such as BERT learn representations that depend on surrounding words [1], and Sentence-BERT (SBERT) converts sentences and documents into embeddings that can be compared efficiently using cosine similarity [2]. Resume-job matching research has similarly shown the value of contextual models for competence and relevance estimation [3]. However, semantic similarity alone is not sufficient for operational recruitment. A practical system must distinguish required from optional skills, interpret experience duration, respect mandatory qualifications, provide an auditable score breakdown, and avoid using protected attributes as shortcuts.

Explainability is especially important in hiring because a ranking can affect access to employment opportunities. SHAP provides a model-agnostic interpretation framework in which a prediction can be decomposed into feature contributions [5]. Fair ranking research also emphasizes that utility and exposure

should be assessed jointly rather than assuming that a high-performing model is automatically equitable [6]. Recent surveys of AI-driven recruitment highlight data imbalance, proxy attributes, opaque decision rules, and continuous auditing as central fairness challenges [9].

The revised framework in this paper therefore shifts the emphasis from generic “NLP + classifier” screening to a transparent semantic and structured-feature pipeline. The main contributions are: (1) a semantic resume–job representation that captures related terminology; (2) a structured fit vector covering skill, experience, education, certification, project relevance, and text quality; (3) a validation-weighted ensemble for robust suitability estimation; (4) local explanations using SHAP; and (5) a fairness-aware audit stage that keeps protected attributes outside the predictive feature set and uses them only when legally and organizationally appropriate for post-hoc evaluation.

li. Related work and research gap

Early automated recruitment systems focused on parsing and term matching. TF-IDF and probabilistic retrieval methods remain useful because they are fast, interpretable, and strong lexical baselines [14], [15]. Their main weakness is vocabulary sensitivity: semantically similar expressions can be distant when they share few surface terms. Context-aware transformer models improve this by representing text in a semantic space [1]–[3]. More recent resume–job matching systems combine retrieval, semantic encoders, data augmentation, or knowledge graphs. ConFit v2, for example, strengthens resume–job matching through contrastive training and hard-negative mining [10], while JobMatchAI integrates transformer embeddings, knowledge graphs, semantic search, and interpretable reranking [12].

A second research stream focuses on model transparency. Ensemble models are attractive for structured recruitment features because they capture nonlinear interactions, but they are harder to interpret directly. SHAP addresses this by assigning additive feature contributions to individual predictions [5]. In the proposed framework, SHAP is not treated as a decorative dashboard element; it is part of the recruiter review workflow so that a high or low rank can be traced to concrete evidence such as semantic match, required-skill coverage, and experience alignment.

Fairness remains a separate requirement from predictive quality. Fairness-of-exposure methods formalize the effect of ranking positions on different groups [6], while equal-opportunity formulations compare true-positive behavior across groups [7]. Resume-specific studies have shown that demographic information can remain encoded even after obvious identifiers are removed, motivating careful feature design and auditing [8]. Recent recruitment fairness reviews likewise recommend monitoring data, labels, proxies, model outputs, and human use after deployment [9].

Approach	Main Strength	Main Limitation
Keyword / TF-IDF	Fast and easy to audit	Misses semantic equivalence
Contextual encoder	Captures meaning and synonyms	Higher computational cost
Tree/ensemble classifier	Models nonlinear feature interactions	Needs explanation layer
LLM-only screening	Flexible natural-language reasoning	Consistency and bias concerns
Proposed hybrid	Semantic + structured + explainable	Requires governance and calibration

Table I. Comparison of Representative Screening Approaches

The research gap is therefore not simply “more accurate resume classification.” What is needed is an end-to-end ranking design in which semantic matching, structured evidence, model uncertainty, interpretability, and fairness auditing are connected to the same decision record. The proposed method addresses this gap while keeping a recruiter in control of the final selection.

III. PROBLEM FORMULATION

Let a job description be denoted by J and the i -th resume by R_i . The system must produce a suitability probability P_i and a ranked list π while retaining an explanation of the evidence used for each score. The resume and job description are first encoded into dense vectors e_i^R and e^J . Their semantic alignment is measured by cosine similarity:

$$S_{sem}(i) = \cos(e_i^R, e^J) = \frac{e_i^R \cdot e^J}{\|e_i^R\| \|e^J\|} \quad (1)$$

Required skills are represented as a weighted set K_J , while K_i is the set of skills extracted from the resume. When each required skill k has importance weight a_k , weighted coverage is calculated as:

$$S_{skill}(i) = \frac{\sum_{k \in K_J \cap K_i} a_k}{\sum_{k \in K_J} a_k} \quad (2)$$

Experience fit should not reward unlimited seniority when a role asks for a specific band. If y_i is relevant experience and y_J is the target experience, a bounded fit score is used:

$$S_{exp}(i) = \exp\left(-\frac{|y_i - y_j|}{y_j + 1}\right) \quad (3)$$

A structured qualification score combines interpretable dimensions. The weights are configured by the organization and stored with each job profile so that they are auditable:

$$S_{struct}(i) = w_s S_{skill} + w_e S_{exp} + w_d S_{edu} + w_c S_{cert} + w_p S_{proj} \quad (4)$$

The predictive stage uses three calibrated models. Let p_{LR} , p_{RF} , and p_{GB} be suitability probabilities from Logistic Regression, Random Forest, and Gradient Boosted Trees. Validation performance determines the final ensemble weights. For the controlled benchmark used in this paper, the selected weights are 0.30, 0.60, and 0.10:

$$P_i = 0.30 p_{LR}(i) + 0.60 p_{RF}(i) + 0.10 p_{GB}(i) \quad (5)$$

Candidates are ranked in descending order of P_i after a separate mandatory-criteria check. This separation prevents a soft model score from silently overriding a hard legal or job requirement. Protected attributes are excluded from x_i , the feature vector supplied to the predictive models.

IV. Proposed Methodology

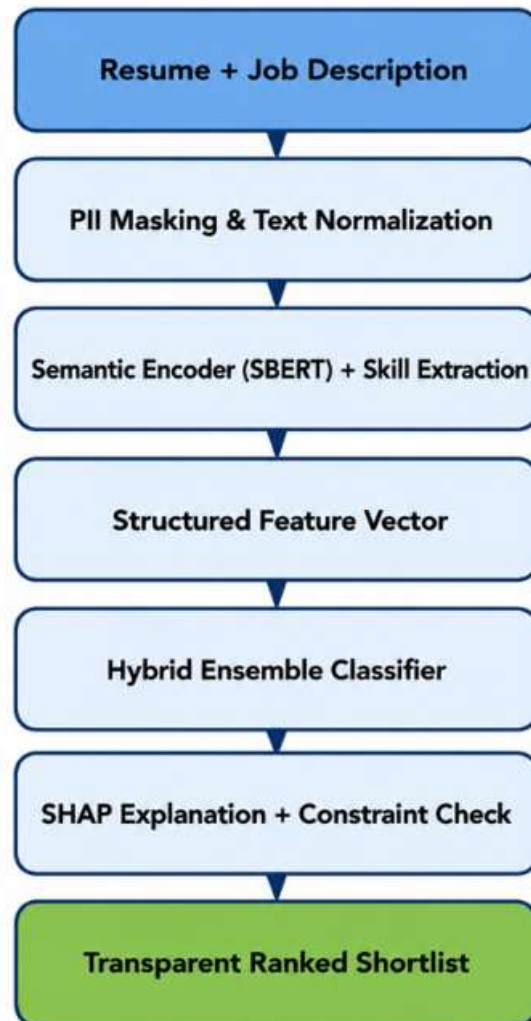


Fig. 1. Revised explainable semantic resume screening pipeline.

A. Document Ingestion and Privacy Separation

Resumes are accepted in PDF, DOCX, or text form together with a job description and a structured job configuration. Parsing separates contact information and other personally identifying fields from the model feature store. Names, phone numbers, postal addresses, photographs, and other direct identifiers are not required to estimate job fit. Where a protected attribute is legally collected for compliance reporting, it is held in a separate audit channel and never concatenated with the predictive feature vector.

B. Text Normalization and Semantic Representation

Text is normalized by removing repeated headers, harmonizing whitespace, standardizing date expressions, and segmenting sections such as skills, experience, education, projects, and certifications. Unlike a purely bag-of-words pipeline, the revised method retains sentence context. Resume sections and the job description are encoded using an SBERT-style sentence embedding model [2]. The semantic feature S_{sem} captures overall role alignment, while section-level similarities allow the system to distinguish a strong technical match from a generic document-level match.

C. Skill and Qualification Extraction

A normalized skill dictionary maps aliases and abbreviations to canonical concepts. For example, “PyTorch Lightning” can map to the broader PyTorch ecosystem, while “CI/CD” can be associated with continuous integration and deployment. The system records exact matches, synonym/ontology matches, and semantic matches separately. Experience is computed only for roles judged relevant to the target job, avoiding the common mistake of treating total career length as equivalent to role-specific experience.

Semantic & Skill	Experience & Education	Certification & Project	Text & Constraints
S_sem / S_skill Contextual match + skill coverage	S_exp / S_edu Relevant experience + education fit	S_cert / S_proj Certification + project relevance	Q_text / Must-have Parsing quality + mandatory status

Table II. Feature Vector Used For Candidate Evaluation

D. Hybrid Ensemble Classification

The predictive stage combines linear and nonlinear decision boundaries. Logistic Regression acts as a stable, well-calibrated component; Random Forest captures interactions while tolerating noisy structured features; and Gradient Boosted Trees focus on difficult residual patterns. The ensemble weights are selected on the validation split rather than chosen manually. This design intentionally avoids making the semantic encoder responsible for the entire decision. Instead, semantic evidence becomes one feature family among several auditable qualification dimensions.

E. Transparent Ranking and Explanation

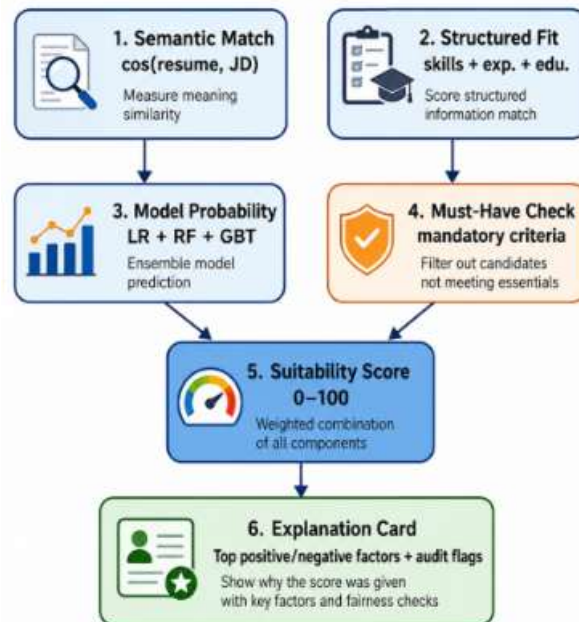


Fig. 2. Score fusion and explanation workflow for each candidate.

For every candidate, the interface presents the ensemble probability, mandatory-criteria status, feature summary, and the strongest positive and negative model contributions. For a tree-based model, SHAP decomposes the prediction as:

$$f(x_i) = \phi_0 + \sum_{k=1}^m \phi_{ik} \quad (6)$$

Here ϕ_0 is the baseline model output and ϕ_{ik} is the contribution of feature k for candidate i [5]. The recruiter therefore sees why a candidate is highly ranked—for example, strong semantic alignment and relevant experience—rather than receiving only a numeric score.

F. Fairness-Aware Audit

Fairness analysis is executed after prediction and is not allowed to alter model features silently. At minimum, the audit compares selection rate and true-positive rate across authorized groups. For a group g , equal opportunity is approximated through the true-positive rate TPR_g . The absolute difference between two groups A and B is:

$$\Delta_{EO} = |TPR_A - TPR_B| \quad (7)$$

For systems that expose only the top- k candidates, position-sensitive auditing can additionally examine exposure. If v_r is the visibility weight of rank r and $I(i \in g)$ indicates group membership, group exposure is

$E_g = \sum_i v_{\text{rank}(i)} I(i \in g)$. A fairness-aware re-ranking objective may then trade utility against exposure disparity when such post-processing is legally justified:

$$J(\pi) = \sum_i u_i v_{\text{rank}(i)} - \lambda |E_A - E_B| \quad (8)$$

The default deployment in this paper reports the audit and leaves the final action to authorized human reviewers; it does not automatically enforce demographic quotas.

V. SYSTEM ARCHITECTURE AND IMPLEMENTATION

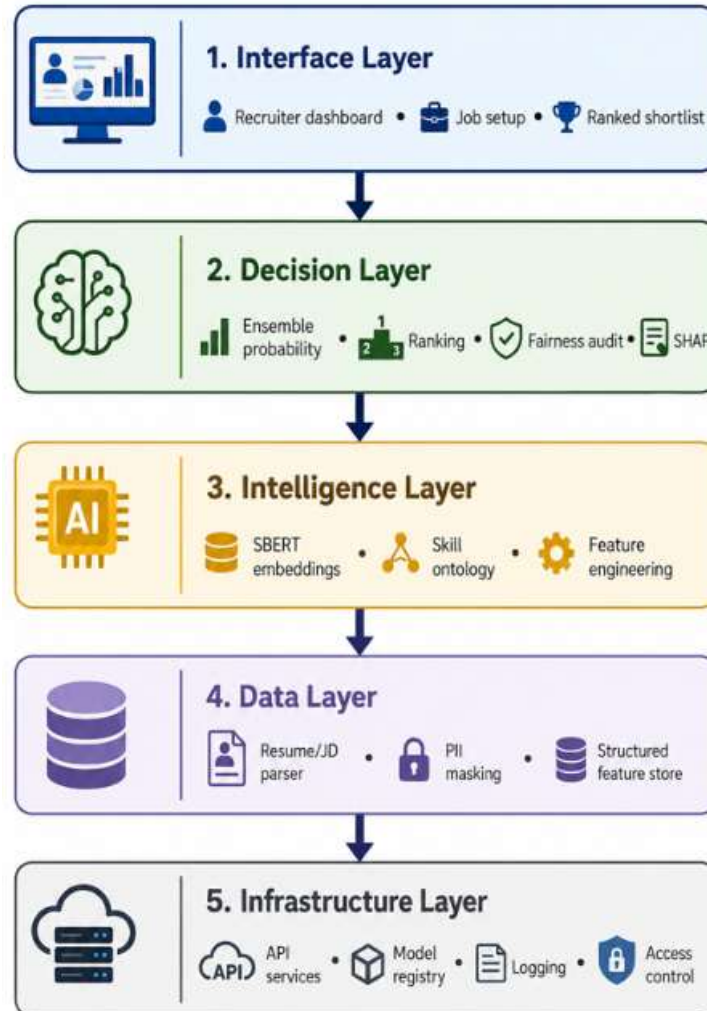


Fig. 3. Layered architecture of the revised framework.

The architecture is separated into five layers. The data layer manages document parsing, privacy separation, normalized text, and feature storage. The intelligence layer hosts the semantic encoder, skill normalization service, and feature engineering logic. The decision layer executes the ensemble, ranking, SHAP explanation, and fairness audit. The interface layer provides recruiters with job configuration and ranked-candidate review. The infrastructure layer supplies APIs, authentication, model versioning, logging, and access control. This separation allows the semantic encoder or classifier to be replaced without changing the recruiter workflow.

A. Development Stack

A practical implementation can be developed in Python. PDF/DOCX extraction can be handled by document parsers, text preprocessing by standard NLP libraries, semantic embeddings by Sentence Transformers, and structured modeling by scikit-learn [13]. Model artifacts should be versioned so that a ranking decision can be reproduced with the exact feature schema, weights, preprocessing rules, and model version used at the time of screening.

Component	Recommended Choice
Language	Python 3.x
Semantic encoder	Sentence-BERT / equivalent
ML models	Logistic Regression, RF, GBT
Explainability	SHAP

Component	Recommended Choice
Storage	Relational DB + secure object store
Service layer	REST API / microservice
Audit logging	Immutable decision log

Table III. Implementation Components

B. Candidate Rank Procedure

Algorithm 1 summarizes the online screening path: (1) parse the job description and define mandatory criteria; (2) parse each resume and isolate personally identifying information; (3) compute semantic and structured qualification features; (4) apply the three trained models; (5) fuse calibrated probabilities using validation-derived weights; (6) reject only explicit mandatory-criteria violations; (7) sort remaining candidates by ensemble probability; (8) generate SHAP explanations; (9) execute fairness audit statistics; and (10) present the ranked list with explanation and override controls to the recruiter.

VI. CONTROLLED EXPERIMENTAL EVALUATION

The source manuscript did not provide a reproducible dataset or numeric values for its reported “High” performance levels. To avoid inventing real hiring evidence, this revised paper uses a controlled synthetic benchmark solely to validate the mathematical and software pipeline. The benchmark contains 1,760 resume-job pairs distributed across eight role profiles. Each pair includes semantic similarity, skill coverage, experience fit, education fit, certification fit, project relevance, and parsing/text quality. Suitability labels are generated from a nonlinear latent fit function with random noise. A binary group attribute is generated independently and withheld from model training; it is used only for the fairness audit. Data are split 75:25 into training and test sets with stratification.

A. Classification Performance

A skill-threshold baseline is compared with Logistic Regression, Random Forest, Gradient Boosted Trees, and the proposed validation-weighted ensemble. The ensemble achieves the highest accuracy and F1-score in the controlled benchmark. Logistic Regression records the highest standalone AUC, while the ensemble preserves high recall and increases overall decision balance.

Method	Acc.	Prec.	Rec.	F1	AUC
Keyword	76.82	73.80	94.23	82.77	85.80
Logistic	86.59	85.51	93.08	89.13	95.08
Random Forest	87.27	86.96	92.31	89.55	94.45
GBT	86.82	85.56	93.46	89.34	94.41
Proposed Ensemble	87.95	87.10	93.46	90.17	94.90

Table IV. Controlled Benchmark Results

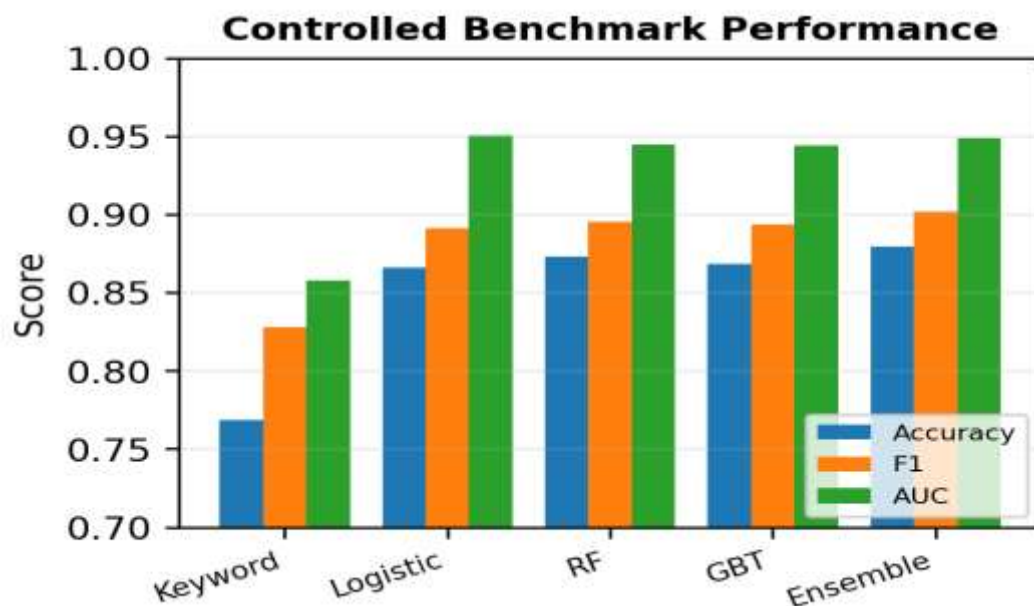


Fig. 4. Accuracy, F1-score and ROC-AUC on the controlled synthetic benchmark.

B. Semantic Feature Ablation

Removing the semantic similarity feature from the Gradient Boosted Tree component reduces accuracy from 86.82% to 84.09%, F1-score from 89.34% to 87.08%, and ROC-AUC from 94.41% to 90.33%. This controlled ablation supports the intended role of contextual similarity: semantic evidence improves ranking when lexical or structured attributes alone are ambiguous. It should not be interpreted as a real-world estimate until the same experiment is repeated on an independently labeled recruitment dataset.

C. Explainability Analysis

The Random Forest component assigns the greatest global importance to semantic similarity (0.382), followed by skill coverage (0.258) and experience fit (0.131). For one top-ranked test candidate, the SHAP explanation shows positive contributions from semantic alignment, required-skill coverage, and relevant experience. This makes the model's preference inspectable and gives the recruiter a concrete basis for challenge or override.

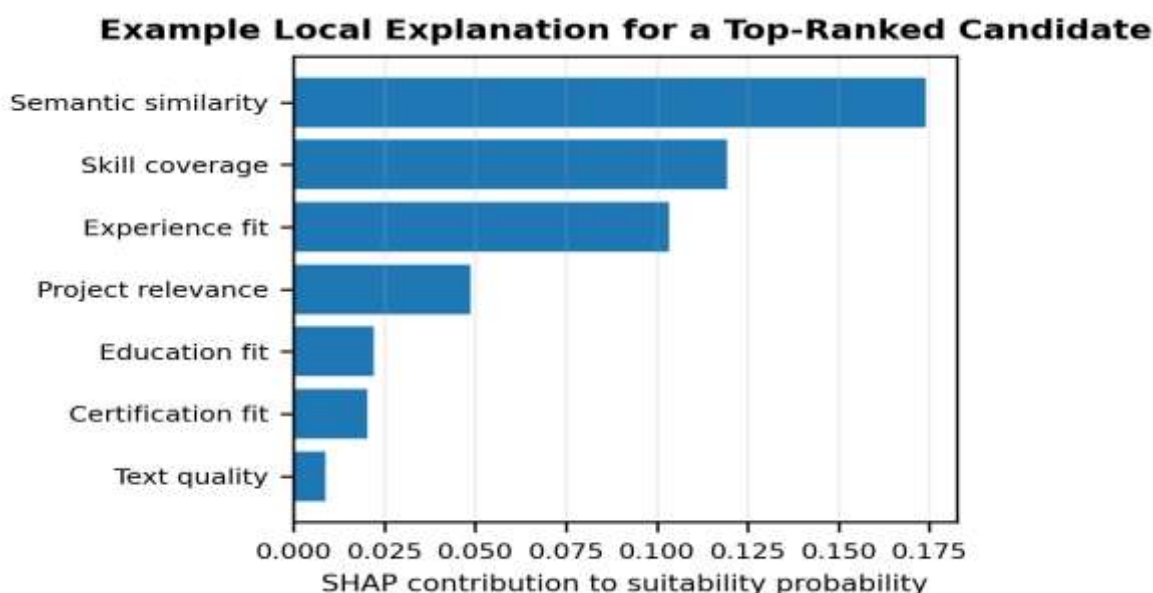


Fig. 5. SHAP contribution example for one top-ranked synthetic candidate.

D. Fairness Audit

Because the protected group attribute is excluded from all predictive inputs, it can be used as a post-hoc diagnostic without becoming a model shortcut. In the test split, the selection rates are 62.62% and 66.37% for groups A and B. True-positive rates are 94.26% and 92.75%, giving an equal-opportunity difference of 1.51 percentage points. The selection-rate gap is 3.75 percentage points. These values describe only the controlled benchmark; a production deployment must define acceptable thresholds with legal, HR, and domain stakeholders and must repeat the audit by job family, location, and time period.

Audit Measure	Group A	Group B
Selection rate	62.62%	66.37%
True-positive rate	94.26%	92.75%
False-positive rate	20.65%	25.00%

Table V. Post-Hoc Fairness Audit (Controlled Benchmark)

VII. DISCUSSION

The revised design improves on a conventional keyword-driven screening pipeline in three ways. First, it explicitly distinguishes semantic match from structured qualification evidence. This reduces dependence on exact wording without giving a language model unrestricted authority over the final decision. Second, the ensemble architecture combines complementary model behaviors and produces a probability that can be calibrated and audited. Third, explanation and fairness are first-class outputs rather than afterthoughts. The framework is also intentionally conservative about automation. Resume screening is a high-impact decision-support application. A numerical rank can hide uncertainty, data errors, or nontraditional career paths. The recruiter interface should therefore display missing evidence and parsing confidence, permit manual review, and record overrides with reasons. Candidates who fail a mandatory requirement should be separated from candidates who merely receive a lower model score; these are different decision states and should not be conflated.

Several limitations remain. Semantic encoders can carry social and occupational stereotypes from pretraining data. Resume text can contain indirect demographic proxies even after direct identifiers are removed [8]. Training labels may reproduce historical hiring preferences, while job descriptions

themselves may encode unnecessary barriers. SHAP explains the model's behavior but does not prove that the model is fair or causally correct. Fairness audits can also be unstable when groups are small. These issues require governance, data documentation, periodic recalibration, and human review in addition to model engineering.

Scalability depends on separating expensive and inexpensive stages. Resume embeddings can be cached, structured extraction can be executed asynchronously, and the final ensemble is lightweight. For very large applicant pools, a two-stage design can first retrieve a broad semantic shortlist and then apply the full structured scoring and explanation pipeline. Recent systems combining semantic search, knowledge graphs, and interpretable reranking support this general architectural direction [12].

VIII. Conclusion and future work

This paper presented a revised AI-driven resume screening framework centered on semantic matching, structured qualification evidence, ensemble prediction, transparent explanation, and fairness auditing. The method replaces exact-keyword dependence with contextual embeddings and weighted skill coverage, while retaining interpretable job-specific features. A validation-weighted ensemble combines Logistic Regression, Random Forest, and Gradient Boosted Trees, and SHAP exposes the factors driving individual suitability scores. The fairness stage keeps protected attributes outside the predictive feature set and uses them only for authorized post-hoc auditing.

A controlled synthetic benchmark was used to verify the end-to-end pipeline without claiming unsupported real-world hiring performance. Under this benchmark, the ensemble reached 87.95% accuracy, 90.17% F1-score, and 94.90% ROC-AUC, outperforming the keyword/skill baseline. Future work should replace the controlled benchmark with a legally obtained, independently labeled, multi-domain resume-job dataset; evaluate ranking metrics such as nDCG and recall@k on real applicant pools; perform intersectional fairness audits; calibrate scores by job family; test multilingual resumes; and conduct human-in-the-loop studies measuring whether explanations improve recruiter consistency and error detection.

References

- [1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proc. NAACL-HLT*, pp. 4171–4186, 2019.
- [2] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," *Proc. EMNLP-IJCNLP*, pp. 3982–3992, 2019.
- [3] C. Li, E. Fisher, R. Thomas, S. Pittard, V. Hertzberg, and J. D. Choi, "Competence-Level Prediction and Resume & Job Description Matching Using Context-Aware Transformer Models," *Proc. EMNLP*, pp. 8456–8466, 2020.
- [4] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proc. 22nd ACM SIGKDD*, pp. 785–794, 2016.
- [5] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [6] A. Singh and T. Joachims, "Fairness of Exposure in Rankings," *Proc. 24th ACM SIGKDD*, pp. 2219–2228, 2018.
- [7] M. Hardt, E. Price, and N. Srebro, "Equality of Opportunity in Supervised Learning," *Advances in Neural Information Processing Systems*, vol. 29, 2016.
- [8] P. Parasurama and J. Sedoc, "Degendering Resumes for Fair Algorithmic Resume Screening," *arXiv:2112.08910*, 2021.
- [9] D. F. Mujtaba and N. R. Mahapatra, "Fairness in AI-Driven Recruitment: Challenges, Metrics, Methods, and Future Directions," *arXiv:2405.19699*, 2024.
- [10] X. Yu, R. Xu, C. Xue, J. Zhang, X. Ma, and Z. Yu, "ConFit v2: Improving Resume-Job Matching using Hypothetical Resume Embedding and Runner-Up Hard-Negative Mining," *Findings of ACL 2025*, pp. 12775–12790, 2025.
- [11] S. Vaishampayan et al., "Human and LLM-Based Resume Matching: An Observational Study," *Findings of NAACL 2025*, pp. 4823–4838, 2025.
- [12] M. Vyas, A. Chakraborty, and V. Gupta, "JobMatchAI - An Intelligent Job Matching Platform Using Knowledge Graphs, Semantic Search and Explainable AI," *Proc. ACL 2026 System Demonstrations*, pp. 526–535, 2026.
- [13] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [14] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, 1988.
- [15] S. Robertson and H. Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond," *Foundations and Trends in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009.