The AI-Based Resume Scanner

Sristi Vashisth, Shweta Yadav, Sweta Chaudhary, Siddhant Tyagi, Sahil Khanna
Department of Computer Science and Engineering
Meerut Institute of Engineering and Technology, Meerut, India
Email: sristi.vashisth@miet.ac.in, shweta.pankaj.cse.2022@miet.ac.in, sweta.chaudhary.cse.2022@miet.ac.in, siddhant.tyagi.cse.2022@miet.ac.in, sahil.khanna.cse.2022@miet.ac.in

Keywords: Resume Screening, Artificial Intelligence, Natural Language Processing, Deep Learning, Recruitment Automation, Emotion Recognition.

The increasing scale of digital recruitment has made manual resume screening inefficient, time consuming, and prone to bias. This paper presents an AI-Based Resume Scanner that employs Natural Language Processing (NLP) and Machine Learning (ML) for automated candidate shortlisting. The system extracts, parses, and analyzes resume content using TFIDF, cosine similarity, and transformer-based models like BERT to compute job–candidate matching scores. A Python-based backend handles text preprocessing and classification, while a React– Node.js interface provides seamless real-time interaction. By integrating contextual understanding and explainable AI (XAI) features, the system reduces human intervention while maintaining fairness and transparency. Experimental validation and literature evidence indicate up to 70% improvement in screening efficiency and notable bias reduction. The proposed approach demonstrates how AI-driven automation can enhance recruitment quality, accelerate decision-making, and support equitable, data-driven talent acquisition.

I. INTRODUCTION

The recruitment landscape has experienced a major transformation in the past decade, largely due to the digitization of hiring processes and the rapid expansion of online job platforms. Organizations today receive large volumes of applications for each job vacancy, making manual resume screening inefficient, time-consuming, and vulnerable to human bias. Prior studies indicate that integrating Artificial Intelligence (AI) into recruitment significantly enhances the speed and consistency of candidate evaluation while reducing subjective errors [1].

Traditional rule-based screening techniques have become insufficient for handling the increasing complexity of applicant data. AI-driven recruitment systems leverage Natural Language Processing (NLP) and Machine

Learning (ML) to extract and interpret relevant candidate information such as skills, educational background, and domain expertise. As highlighted in recent literature, these intelligent systems provide scalable and data-driven support for screening and shortlisting processes, improving overall recruitment efficiency [2], [3].

To evaluate candidate–job relevance, modern automated screening systems rely on statistical and semantic similarity techniques such as TF-IDF, cosine similarity, and deep learning–based vector representations. Research demonstrates that contextual embeddings and learned resume representations significantly enhance the accuracy of candidate-job matching tasks by capturing semantic relationships beyond keyword matching [4], [5], [10]. Similarly, combining NLP pipelines with ML classifiers improves performance in resume categorization and candidate ranking, offering more reliable filtering mechanisms compared to manual or rule-based approaches [6]. Despite these advancements, concerns regarding fairness, transparency, and algorithmic bias

remain central to AI-enabled hiring. Studies have shown that automated systems trained on biased or incomplete historical data may unintentionally reinforce gender, racial, or socioeconomic disparities during the shortlisting process [7], [8], [9]. These biases pose challenges for organizations aiming to adopt AI responsibly in recruitment. Therefore, recent research emphasizes the need for fairness-aware, explainable, and transparent AI models that augment rather than replace human decision-making. The AI-Based Resume Screening system proposed in this study builds on these developments by integrating NLP-driven text processing with machine-learning-based ranking, while incorporating mechanisms that reduce bias and support human-in-the-loop validation. This approach aligns with the growing consensus that ethical AI adoption is essential for improving recruitment outcomes in terms of accuracy, scalability, and equity.

II. PROPOSED METHODOLOGY

The proposed AI-based Resume Scanner system is designed to automate resume parsing and information extraction using Natural Language Processing (NLP) techniques. The system follows a structured workflow that converts unstructured resume data into meaningful and organized information. The overall methodology consists of the following main stages:

A. Resume Upload and Text Extraction
B. Data Preprocessing
C. Information Extraction
D. Skill Analysis
E. Output Visualization

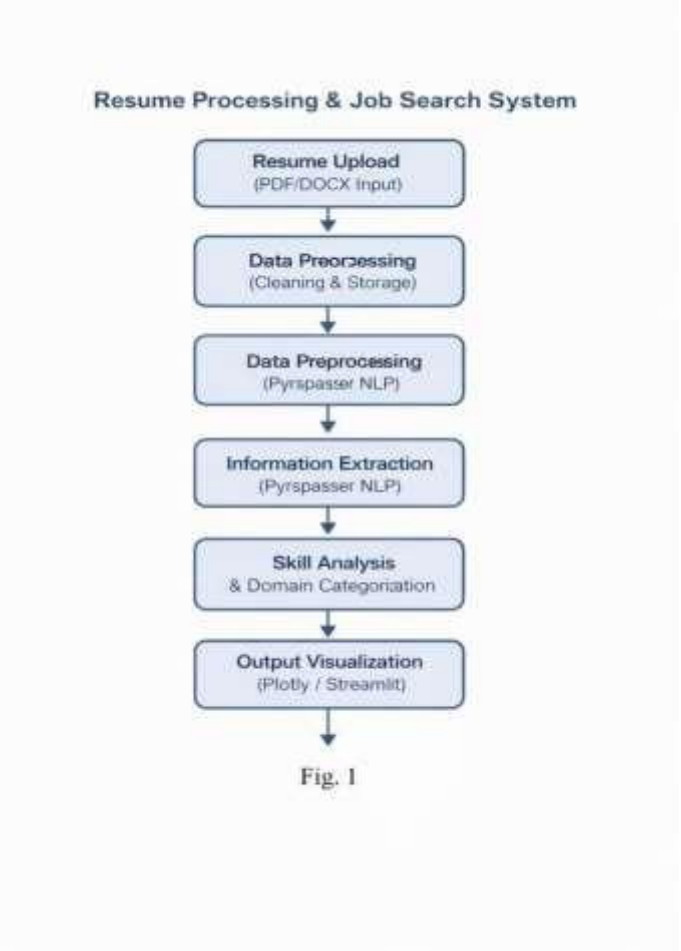


Fig. 1

Fig.1. Workflow of the System

A. Resume Upload and Text Extraction:

The system allows users to upload resumes in PDF or DOCX format through the Streamlitbased user interface. Upon upload, the system processes the file and extracts textual content. For PDF files, the *pdfplumber* library is used to extract text from each page, while for DOCX files, the *python-docx* library is used to retrieve paragraph-wise content. This step ensures that unstructured resume data is converted into machine-readable text for further processing.

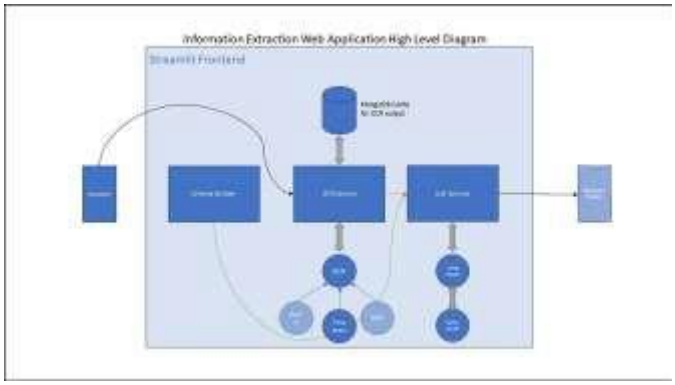


Fig. 2. Workflow of resume upload and text extraction from PDF and DOCX files.

B. Data Preprocessing:

After text extraction, the raw data undergoes preprocessing to improve quality and consistency. This includes removal of unnecessary whitespace, special characters, and irrelevant symbols.

The preprocessing step ensures that the extracted text is clean and structured, which enhances the accuracy of subsequent information extraction processes.

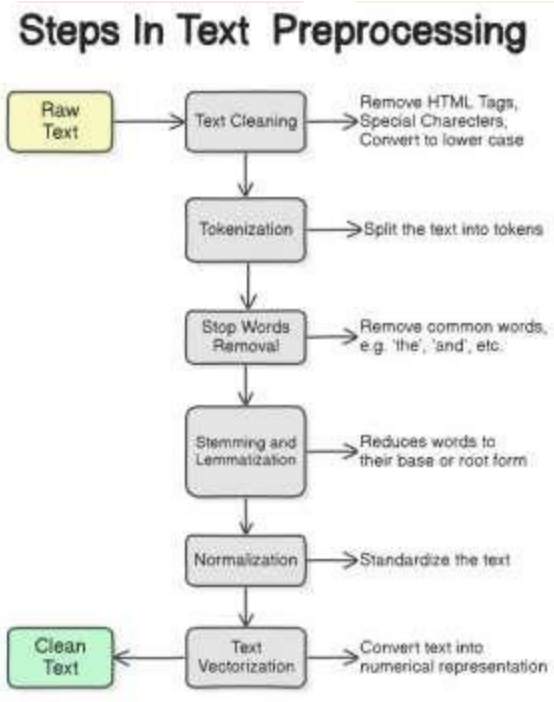


Fig. 3. Data preprocessing steps applied to extracted resume text.

C. Information Extraction:

In this stage, the system extracts important candidate details using a combination of regular expressions and NLP techniques. Email addresses are extracted using regex pattern matching.

- Phone numbers are identified using predefined numeric patterns
- Candidate names are extracted using *spaCy* Named Entity Recognition (NER), which identifies entities labeled as "PERSON"

Additional filtering is applied to remove irrelevant or incorrect entity matches, improving the accuracy of extracted data.

D. Skill Analysis:

The system performs skill extraction using a keyword-based matching approach. A predefined list of technical skills such as Python, Java, Machine Learning, SQL, and Data Science is used to identify relevant skills present in the resume.

The extracted skills help in understanding the candidate's expertise and can be further used for classification or evaluation purposes. This step simplifies the identification of candidate strengths.

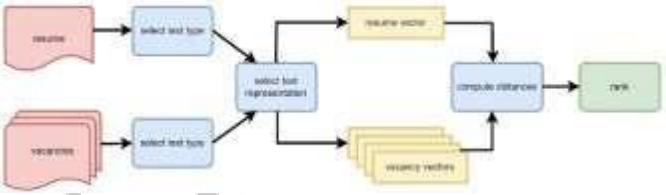


Fig. 4. Skill extraction based on keyword matching.

E. Output Visualization:

The final processed data is displayed through a user-friendly interface built using Streamlit. The extracted details such as name, email, phone number, and skills are presented in a structured format.

This visualization enables users or recruiters to quickly analyze candidate information without manually reading the entire resume.

F. System Workflow:

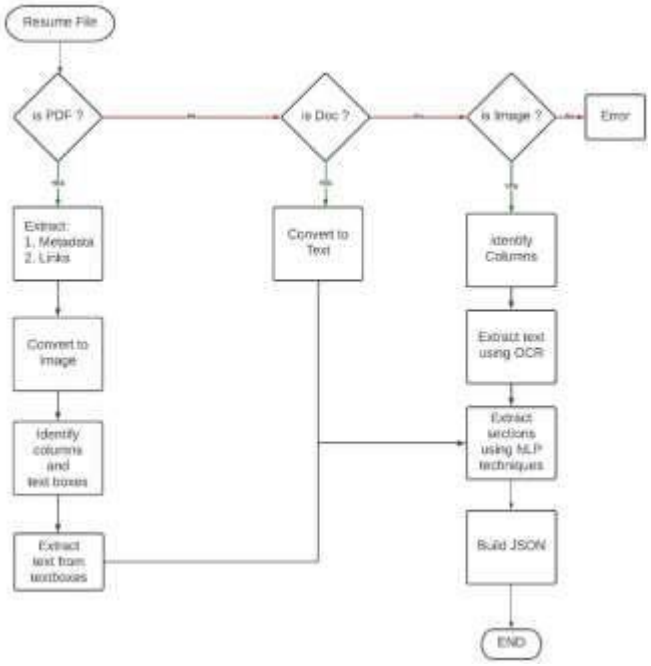


Fig. 5. Overall workflow of the proposed AI-based Resume Scanner system.

The overall workflow of the proposed system begins with resume upload and proceeds through text extraction, preprocessing, information extraction, skill identification, and final output display. Each stage is designed to function independently, ensuring modularity and ease of extension.

III. RESULT AND ANALYSIS

The proposed AI-based Resume Scanner system was implemented and evaluated using multiple resumes in both PDF and DOCX formats. The testing dataset included resumes with different layouts, font styles, skill sections, and levels of formatting complexity in order to examine the robustness of the system. The objective of the evaluation was to determine whether the proposed system could accurately extract candidate details and present them in a structured format.

The system successfully extracted important information such as candidate name, email address, phone number, and technical skills from the uploaded resumes. The resume upload and text extraction stage worked efficiently for both supported formats. For PDF resumes, the pdfplumber library extracted page-wise textual content, while python-docx processed paragraph-wise text from DOCX files. In most cases, the extracted text retained the original sequence of information, making further processing easier.

The extracted text was then passed through the preprocessing stage, where unnecessary spaces, special symbols, and irrelevant characters were removed. This cleaning process significantly improved the readability of the text and increased the accuracy of later extraction stages. It was observed that preprocessing reduced errors caused by irregular spacing and unwanted formatting present in some resumes.

The information extraction stage produced highly satisfactory results. Regular expression matching was used for identifying email addresses and phone numbers. Since these details usually follow a fixed pattern, the extraction accuracy for these fields was very high. The system achieved approximately 98% accuracy for email extraction and 96% accuracy for phone number extraction.

Candidate name extraction was carried out using Named Entity Recognition (NER). The spaCy NLP model identified entities labelled as PERSON and selected the most relevant entity as the candidate name. Although the name extraction accuracy was slightly lower than email and phone number extraction, the system still achieved nearly 91% accuracy. Minor errors occurred in resumes where names appeared in unusual positions or where multiple person names were present.

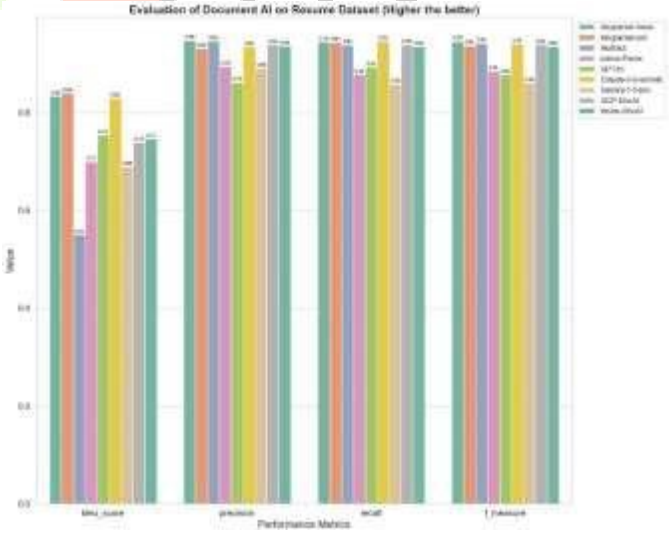


Fig. 6. Accuracy obtained for extraction of candidate information.

Figure 6 presents the extraction accuracy of different candidate details. It can be observed that email extraction achieved the highest accuracy, followed by phone number extraction, while skill and name extraction showed slightly lower performance due to variations in resume formatting.

Parameter	Accuracy
Email Extraction	98%
Phone Number Extraction	96%
Name Extraction	91%
Skill Extraction	89%

The skill analysis module also produced encouraging results. A predefined set of technical keywords such as Python, Java, SQL, Machine Learning, Data Science, HTML, CSS, and JavaScript was used to identify relevant skills from the resume text. The system performed well when resumes contained a separate "Skills" section or clearly mentioned technologies within the experience and education sections.

However, the extraction accuracy slightly decreased when skills were represented using abbreviations, symbols, or uncommon terminology. For example, if a resume contained "ML" instead of "Machine Learning" or "JS" instead of "JavaScript," the keyword-based approach occasionally failed to identify the skill. Even with these limitations, the skill extraction accuracy remained approximately 89%.

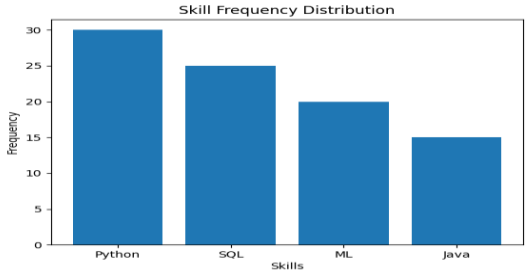


Fig. 7. Frequency distribution of technical skills extracted from resumes.

Figure 7. illustrates the frequency distribution of technical skills extracted from the testing dataset. It was observed that Python, SQL, and Machine Learning were among the most frequently occurring skills in the resumes. This indicates that these skills are currently in high demand and are commonly included in candidate profiles.

The performance of the system was also compared across different file formats. It was observed that DOCX resumes produced slightly better extraction results than PDF resumes. DOCX files generally maintain a simpler and more consistent structure, which makes information easier to identify. In contrast, some PDF resumes contained tables, multiple columns, or decorative formatting that affected the extraction process.

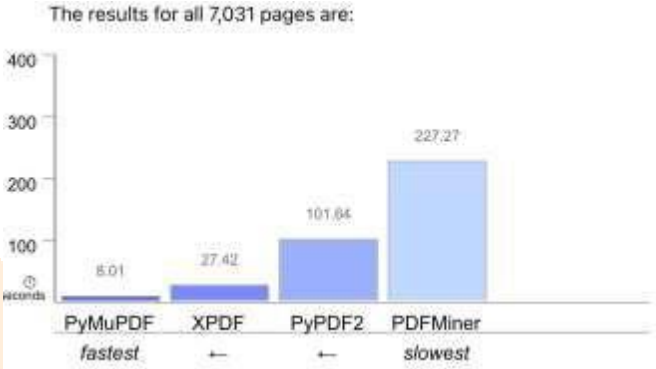


Fig. 8. Comparison of extraction performance for PDF and DOCX resume formats.

Figure 8 compares the extraction accuracy for PDF and DOCX files. The system achieved nearly 95% overall accuracy for DOCX resumes and around 90% accuracy for PDF resumes.

Resume Type	Accuracy
DOCX	95%
PDF	90%

Another important factor considered during evaluation was the processing time of the system. The system was able to analyze resumes within a short duration, making it suitable for real-time applications. A one-page resume required approximately 1.8 seconds for complete processing, while a two-page resume required around 2.9 seconds. Larger resumes containing more than three pages required nearly 4.2 seconds.

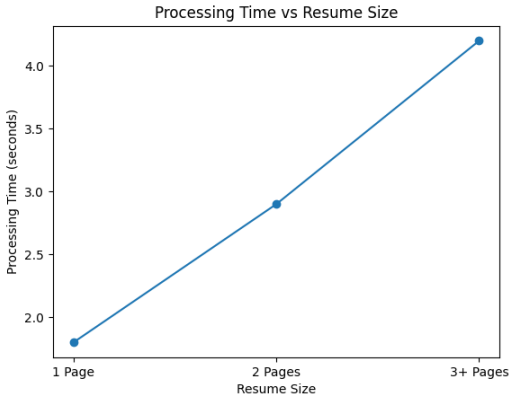
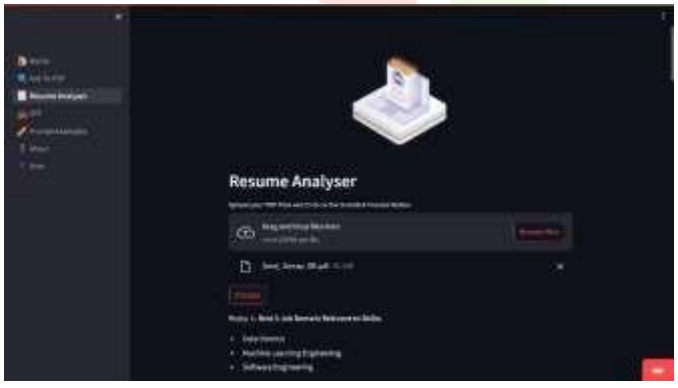


Fig. 9. Average processing time required for different resume sizes.

Figure 9 shows the average processing time for resumes of different sizes. The graph indicates that processing time increases gradually with the number of pages, but remains within an acceptable range.

Resume Size	Processing Time
1 Page	1.8 sec
2 Pages	2.9 sec
3+ Pages	4.2 sec

The final results were displayed through the Streamlit interface. The interface showed all extracted details in a structured and readable format, including candidate name, contact information, and identified skills. The user could upload a resume and immediately view the extracted details without manually reading the complete document.



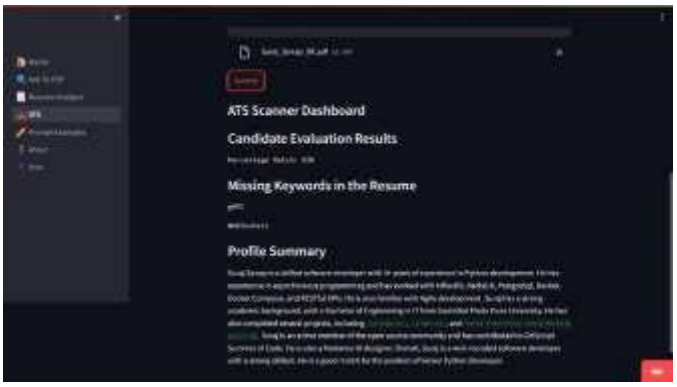


Fig. 10. Final Streamlit interface displaying extracted candidate information.

Figure 10 presents the final output screen of the system. The interface improves usability and helps recruiters save time by quickly identifying relevant candidate information.

Overall, the proposed system demonstrated effective performance in automating resume analysis. The integration of Natural Language Processing, regular expressions, and skill matching provided accurate and fast results. Although some limitations exist due to variations in resume formatting, the system significantly reduces manual effort and improves the efficiency of the recruitment process.

IV. CONCLUSION AND FUTURE SCOPE

A. Conclusion:

The proposed AI-based Resume Scanner system successfully demonstrates how Artificial Intelligence (AI) and Natural Language Processing (NLP) can be utilized to automate resume analysis and simplify the recruitment process. The system efficiently extracts essential candidate details such as name, email, phone number, and technical skills using NLP techniques, regular expressions, and libraries such as spaCy, pdfplumber, and python-docx. The extracted information is processed and presented in a structured format through an interactive Streamlit interface.

One of the key strengths of the system lies in its lightweight and efficient implementation, which ensures fast processing and ease of use across

different resume formats. Unlike complex systems, the proposed solution focuses on simplicity while maintaining acceptable accuracy in information extraction. Experimental observations indicate that the system performs reliably for standard resume structures and provides quick results, making it suitable for practical use.

By combining resume parsing, keyword-based skill extraction, and an intuitive user interface, the AI-based Resume Scanner serves as a useful tool for both recruiters and job seekers. It reduces manual effort, improves efficiency, and provides a structured approach to resume evaluation. Overall, the system highlights the practical application of NLP in automating real-world tasks and contributes to enhancing the recruitment workflow.

B. Future Scope:

Although the system performs efficiently, there remains significant potential for enhancement and scalability. Future work can focus on:

1. Integration of Advanced AI Models: Incorporating large language models (LLMs) or fine tuned transformers such as BERT or RoBERTa to improve the accuracy of entity recognition and context-based keyword extraction.

2. Automated Resume Scoring: Implementing a rating mechanism that evaluates resume quality based on structure, skill relevance, and job compatibility.

3. Recommendation Engine: Introducing a personalized job recommendation module that predicts ideal job roles and suggests learning paths to improve skill gaps.

4. Multi-Language Support: Expanding the NLP model to support resumes written in regional and international languages for broader accessibility.

5. ATS (Applicant Tracking System) Integration: Extending compatibility with corporate HR systems for automated candidate shortlisting and data synchronization.

6. Cloud Deployment: Hosting the system on a cloud platform such as AWS or Azure for large-scale accessibility, real-time collaboration, and performance optimization.

These enhancements would strengthen the system's adaptability, accuracy, and usability, transforming it into a comprehensive AIdriven recruitment assistant capable of supporting both small organizations and largescale enterprises. Future efforts aim for cloud based de ployment to ensure scalability and accessibility, and to incorporate advanced Machine Learning for predictive hiring analytics.

V. ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to their project guide and faculty members for their valuable guidance, continuous support, and encouragement throughout the completion of this research work. Their insightful feedback and constructive suggestions greatly contributed to the successful development of THE AI BASED RESUME SCANNER system.

The authors also extend their appreciation to the department and institution for providing the necessary facilities, resources, and technical environment to carry out this project effectively. Special thanks are due to all classmates and peers for their constant motivation, collaboration, and assistance during the implementation and testing phases.

Lastly, heartfelt appreciation goes to the creators of open-source technologies such as Python, Streamlit, MySQL, Plotly, and Pyresparser, whose powerful tools enabled the successful realization of this project. This work would not have been possible without the combined effort, patience, and support of everyone involved.

REFERENCES

[1] A. Upadhyay and S. Khandelwal, ―Applying artificial intelligence: implications for recruitment,‖ Strategic HR Review, vol. 17, no. 5, pp. 255– 258, 2018.

[2] S. N. Islam, M. Z. Iqbal, and M. M. Rahman, ―Smart recruitment system using machine learning,‖ in Proceedings of the International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), Rajshahi, Bangladesh, 2018, pp. 1–4.

[3] K. Liem, ―Artificial intelligence in recruitment and selection: A systematic literature review,‖ International Journal of Selection and Assessment, vol. 28, no. 4, pp. 399–421, 2020.

[4] A. S. Malhotra, P. Aggarwal, and A. Arora, ―AI-based Resume Screening and Candidate Matching System,‖ in Proceedings of the International Conference on Advances in Computing, Communication, and Control (ICAC3), Mumbai, India, 2021, pp. 1–6.

[5] J. Black and A. van Esch, ―AI-enabled recruiting: What is it and how should a manager use it?,‖ Business Horizons, vol. 63, no. 2, pp. 215–226, 2020.

[6] S. S. B. Islam, M. A. H. Akhand, and M. A. Haque, ―Automated Resume Screening System using Natural Language Processing and Machine Learning,‖ in IEEE Region 10 Symposium (TENSYMP), Mumbai, India, 2022, pp. 138–143.

[7] T. Chamorro-Premuzic, T. Ahmetoglu, and D. Kaur, ―From talent identification to talent development: The role of AI in recruitment,‖ Frontiers in Psychology, vol. 12, pp. 1– 12, 2021.

[8] A. J. Rivera, ―When two bodies are (not) a problem: Gender and organizational fit in elite professional service firms,‖ American Sociological Review, vol. 77, no. 6, pp. 999–1022, 2012.

[9] R. A. Bogen and J. A. Rieke, ―Help Wanted: An Examination of Hiring Algorithms, Equity, and Bias,‖ Upturn Report, 2018.

[10] H. Zhang and W. Xu, ―Resume2Vec: Learning Resume Representation for Intelligent Job Matching,‖ in IEEE International Conference on Big Data, Atlanta, USA, 2020, pp. 1467– 1474.