**Classification and Scoring of Resumes using ML based resume classifier system based on various criterias**

**Dr. Abhishek Goyal[1], Abhishek Verma[2], Alok Ranjan Dubey[3], Anurag kumar[4],**

**Adarsh Mishra[5]**

# [1][2][3][4][5] Computer Science Department, KIET Group of Institutions

# 

**ABSTRACT**

In the modern job market, recruiters face significant challenges in the initial stages of resume screening. Traditional manual resume evaluation is not only time-intensive but also prone to inefficiencies, subjectivity, and human biases. As the number of job applications continues to rise, there is a growing demand for automated solutions to optimize this process. This paper introduces ResuPro, a machine learning (ML)-powered resume classification system designed to automate and enhance resume screening. Utilizing advanced ML algorithms and natural language processing (NLP), ResuPro categorizes resumes into predefined job sectors based on their content. The system efficiently extracts key details such as skills, experience, and company type (Startup, Midlevel, Enterprise), offering recruiters a structured and insightful analysis of applicants. ResuPro classifies resumes by job domain (e.g., IT, Finance, Marketing) and integrates a scoring mechanism to assess resumes based on their relevance to job requirements, including skills, experience, and education. This scoring framework ranks candidates, highlights the most suitable profiles, and provides a data-driven evaluation of their qualifications. The classification model achieves an accuracy of 88%, significantly improving the efficiency of the hiring process. By automating resume analysis, ResuPro minimizes manual workload, accelerates hiring, and enhances decision-making by ensuring objective and unbiased candidate evaluation. This system provides a scalable and reliable approach to resume screening, enabling organizations to streamline recruitment and make fairer, data-driven hiring decisions. This paper discusses the architecture, methodology, and performance of ResuPro, demonstrating how ML and NLP technologies can transform the hiring process, making it more efficient, equitable, and effective.

**Keywords**: Machine Learning, Natural Language Processing, Resume Classification, Recruitment Automation, Resume Scoring

# INTRODUCTION

The recruitment landscape is undergoing a significant transformation, driven by the rapid advancements in technology and the increasing volume of job applications received by organizations. Traditionally, the process of reviewing and screening resumes has been manual, relying heavily on human judgment and the subjective evaluation of recruiters. This standard practice, while effective in smaller scale hiring scenarios, often proves inadequate in today’s dynamic job market where companies receive hundreds, if not thousands, of applications for a single position. The manual review process is time-consuming, resource-intensive, and susceptible to human biases, which can result in overlooking highly qualified candidates or favoring certain profiles based on unconscious biases. [4][6]

As organizations strive to optimize their hiring strategies, the need for automated solutions that can streamline the initial stages of recruitment has become more apparent. The primary challenge lies in effectively filtering through a vast pool of candidates to identify those who best match the job requirements. Traditional keyword-based filtering methods, while faster than manual reviews, often lack the contextual understanding needed to assess the relevance and quality of a resume. Moreover, these systems can be easily manipulated through keyword stuffing, where candidates add multiple buzzwords without possessing the actual skills, leading to inaccurate assessments.

To address these challenges, we propose ResuPro, machine learning (ML)-based solution that aims to revolutionize the way resumes are screened and evaluated. The core objective of ResuPro is to automate the classification and scoring of resumes using state-of-the-art machine learning algorithms and natural language processing (NLP) techniques. By leveraging NLP, the system can extract critical information from resumes, such as educational qualifications, work experience, skills, and company type (Startup, Midlevel, Enterprise). This data is then processed using ML algorithms like Random Forest and Naive Bayes, which categorize the resumes into predefined job domains (e.g., IT, Finance, Marketing) based on their content.

ResuPro introduces a sophisticated scoring mechanism that evaluates resumes against specific job requirements, considering factors such as experience level, skills relevance, and educational background. This approach not only speeds up the recruitment process but also enhances its accuracy by prioritizing candidates whose profiles align closely with the job criteria. The scoring system provides recruiters with a ranked list of candidates, highlighting the most promising ones, thus reducing the time and effort required for initial screenings.

The adoption of machine learning and NLP in resume screening addresses several key pain points in the recruitment process. Firstly, it significantly reduces the time spent on manual reviews, allowing talent acquisition teams to focus on interviewing and engaging with top-tier candidates. Secondly, by analyzing resumes in a consistent and unbiased manner, ResuPro mitigates the risk of human bias, ensuring a fairer evaluation process. Lastly, the scalable nature of this automated system makes it suitable for organizations of all sizes, from startups to large enterprises, looking to enhance their hiring efficiency.[3][5]

In this paper, we explore the design, methodology, and implementation of ResuPro, delving into the various components that make up this ML-based resume classifier. We discuss the technical aspects of NLP preprocessing, feature extraction, and the machine learning models used for classification and scoring. Additionally, we evaluate the performance of the system in real-world scenarios, highlighting its potential to transform the recruitment process by providing a more efficient, accurate, and unbiased method of identifying top talent. Through ResuPro, we aim to contribute a valuable tool to the field of HR technology, addressing the growing demand for automated solutions in talent acquisition.

# OBJECTIVES

The primary objectives of ResuPro are:

1. To provide a user-friendly interface for creating and uploading resumes.
2. To automate the initial screening process of resumes, reducing manual effort and time.
3. To use NLP techniques for extracting key information and understanding the semantic context of resumes.
4. To implement machine learning models for domain classification of resumes.
5. To develop a scoring system that evaluates resume quality based on relevance to job requirements.
6. To enhance decision-making by providing insights into candidate qualifications.

# LITERATURE REVIEW

Resume screening has evolved significantly, transitioning from manual approaches to sophisticated automated systems powered by machine learning (ML) and natural language processing (NLP). Traditional systems primarily relied on keyword matching and basic filtering techniques, which, while straightforward, often failed to capture the deeper contextual meaning of a candidate's qualifications. These limitations led to the exclusion of many qualified candidates whose resumes used unique wording or phrased skills differently. Keyword-based systems lacked the flexibility to account for variations in terminology and failed to analyze the relevance of experience in a broader context, making them less effective for comprehensive candidate evaluations.

Advances in ML and NLP have brought significant improvements to resume screening processes by enabling systems to analyze unstructured data and derive meaningful insights. NLP techniques like Named Entity Recognition (NER) and lemmatization have enhanced the ability to extract and standardize information from resumes, identifying critical details such as job titles, company names, and technical skills. By using lemmatization, systems normalize words to their base forms, ensuring that variations like "programming" and "programmer" are treated equivalently. In addition, vectorization techniques like Term Frequency-Inverse Document Frequency (TF-IDF) allow systems to represent text data in numerical formats, facilitating the application of ML algorithms for classification and analysis. These advancements have shifted resume screening from rigid keyword-based systems to dynamic, context-aware solutions.

The integration of ML models has further strengthened resume screening systems, enabling them to handle high-dimensional data and complex relationships between features. Algorithms like Support Vector Machines (SVM), Random Forest, and K-Nearest Neighbors (KNN) have demonstrated their effectiveness in categorizing resumes based on job requirements. SVM, for instance, is highly effective at finding optimal decision boundaries to separate different classes, while Random Forest uses ensemble learning to reduce overfitting and improve prediction accuracy. These algorithms are well-suited for tasks involving high-dimensional feature spaces, such as resume analysis, where resumes can contain diverse and unstructured information.

Various studies have showcased the effectiveness of these approaches in enhancing recruitment processes. One study proposed an automated resume classification system using ML techniques such as Decision Tree, Random Forest, KNN, and SVM. Using a dataset of 3446 resumes, the system employed NLP for preprocessing and demonstrated high accuracy across all models, emphasizing the potential of ML in automating resume classification and improving recruitment efficiency¹. Another study utilized cosine similarity as a metric for resume classification, combining it with NLP techniques like stop word removal, lemmatization, and TF-IDF vectorization. The system achieved an accuracy of 98.96% using the KNN algorithm, demonstrating the value of semantic understanding in matching resumes to job descriptions[2].

Further exploration of NLP and ML techniques included the development of a model for suggesting resumes based on job roles. This system utilized NER for extracting technical skills, which were then represented using word embeddings. By applying cosine similarity to evaluate matches between resumes and job descriptions, the model achieved an accuracy of 79.8%, showcasing the potential of semantic embeddings in improving resume screening³. Another application focused on a web-based system that ranked resumes on the recruiter side using a scoring mechanism. This system employed SpaCy for NLP preprocessing and demonstrated the practical implementation of automated resume screening in real-world scenarios[14].

The cost-effectiveness of ML-based solutions has also been emphasized in studies aimed at reducing hiring expenses. One such approach used n-grams for feature extraction and various ML classifiers, achieving a maximum accuracy of 78.53% with the Linear Support Vector Machine. This study highlighted the scalability of ML techniques in addressing challenges related to applicant selection and resume analysis⁵. Additionally, a system integrating NLP with ML algorithms like KNN, Random Forest, and SVM demonstrated that Random Forest outperformed others with an accuracy of 94.5%. This research emphasized the ability of ML models to enhance recruitment processes by reducing workload and improving accuracy[6].

The use of ML for ranking resumes has also been investigated, with studies comparing the performance of algorithms like SVM, Naïve Bayes, and KNN. SVM consistently outperformed the other models in terms of accuracy and efficiency, making it a reliable choice for high-dimensional tasks like resume ranking. This approach highlighted the potential of automated systems to save time and resources in recruitment while maintaining accuracy⁷. Another model aimed to predict the likelihood of candidates accepting job offers by analyzing relevant attributes. Using statistical measures and ML algorithms, with Random Forest achieving an accuracy of 94.86%, the study demonstrated the versatility of ML in recruitment, extending beyond resume screening to behavioral predictions[8].

Building on these advancements, systems like ResuPro integrate state-of-the-art NLP techniques and robust ML models to create a comprehensive resume classification solution. By utilizing NER and lemmatization for feature extraction and combining these with models like Naïve Bayes and Random Forest, these systems aim to move beyond keyword matching. They enable a deeper, context-aware analysis of resumes, capturing the relevance of skills and experiences even when expressed in unconventional ways. The hybrid approach allows for scalable processing of large datasets, providing precise candidate evaluations that save both time and resources.

These automated systems demonstrate the transformative potential of ML and NLP in recruitment. By addressing the limitations of traditional keyword-based methods and offering scalable solutions for large-scale recruitment, they have significantly improved the accuracy and efficiency of candidate selection. Moreover, these systems can be tailored to align with specific job requirements, ensuring that the evaluation process is both objective and relevant. As a result, organizations can make more informed hiring decisions, ultimately reducing the time-to-hire and enhancing the overall quality of recruitment outcomes.

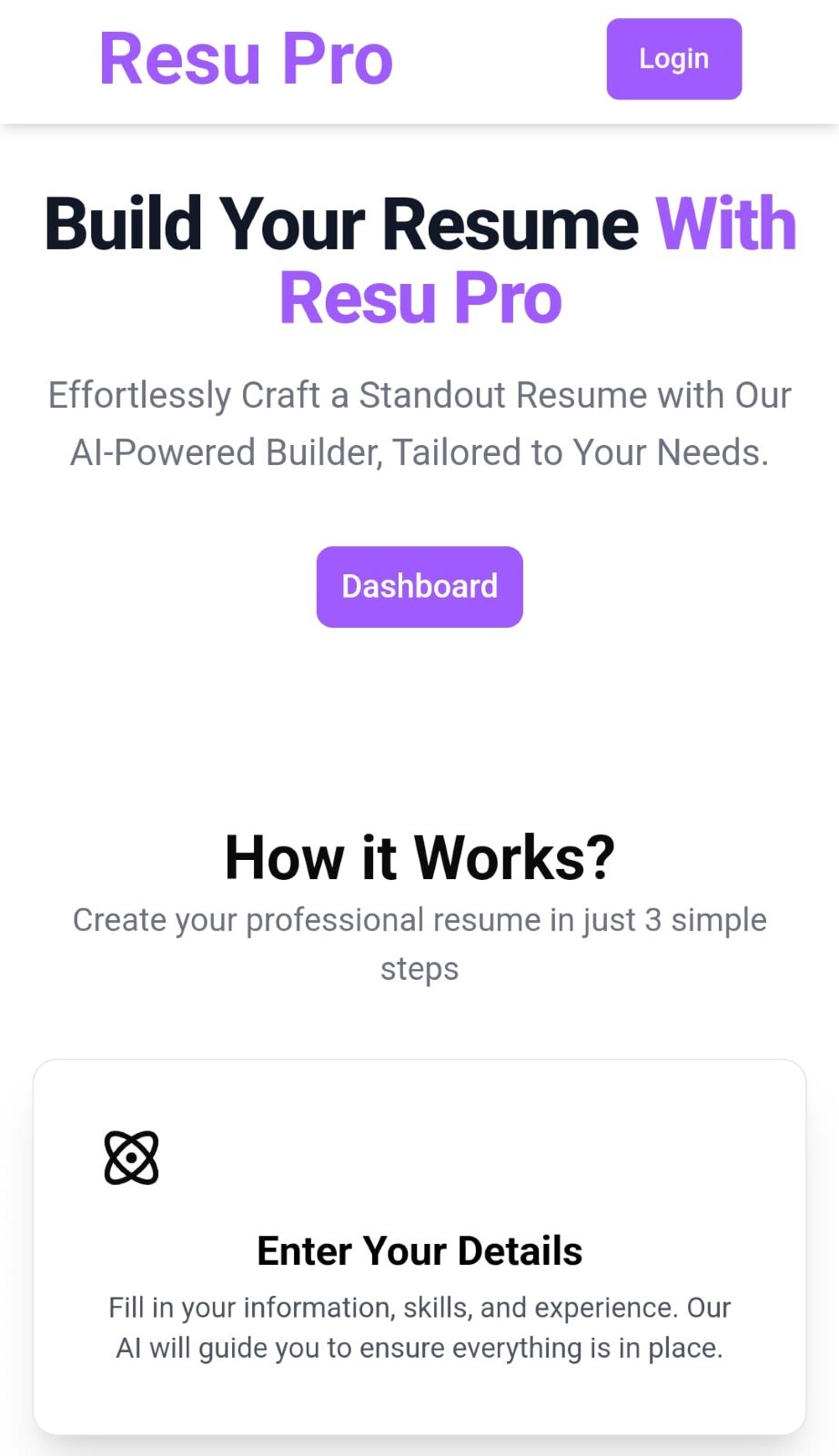
In conclusion, the integration of advanced ML and NLP techniques has revolutionized resume screening, providing intelligent, automated solutions that address the shortcomings of traditional approaches. With continued research and development, these technologies are poised to play an even greater role in shaping the future of recruitment, offering scalable, efficient, and accurate tools for identifying the best candidates in competitive job markets. As more organizations adopt these systems, they can expect to achieve significant time and cost savings, while also improving the quality of their hiring processes.

# PROPOSED METHODOLOGY

**4.1. FRONT-END DEVELOPMENT**

A React-based frontend interface, enhanced with Vite and Tailwind CSS, allows users to create or upload resumes seamlessly.

React.js enables a single-page application (SPA) that allows for real-time updates to the UI without requiring full page reloads. Key features implemented include:



The UI guides candidates through the resume creation process with prompts and suggestions, ensuring a structured and comprehensive resume format.

**4.2. BACK-END DEVELOPMENT**

The back-end API is built using Node.js and Express.js, with the following core functionalities:

**Authentication and Authorization:** User authentication is handled using JSON Web Tokens (JWT), providing secure access to documents according to assigned user roles.

**CRUD Operations:** Users can upload new documents, update metadata, and delete files based on their role.

**4.3. DATABASE MANAGEMENT**

The use of MongoDB as a NoSQL database enhances the flexibility and scalability of the ResuPro system. MongoDB’s ability to store dynamic, unstructured data makes it an ideal choice for managing various types of resume documents and related metadata.

In ResuPro, MongoDB not only stores the actual resume documents but also stores critical metadata, such as:

**User Data:** To keep the records of the user’s resume, and along with his personal information.

**Upload Dates:** To track when resumes were submitted.

**Authors:** Information about the users submitting the resumes.

**Tags:** Keywords or categories relevant to the job domain (e.g., "Software Engineer," "Finance").

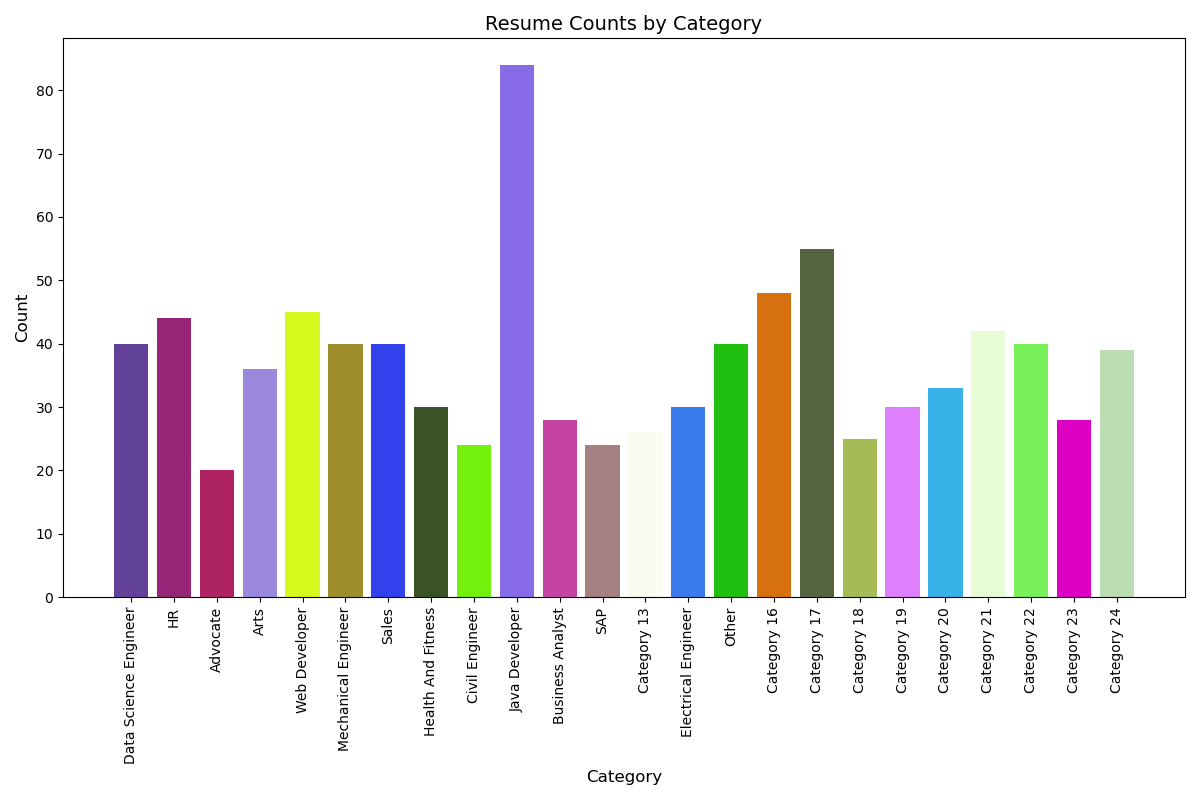
MongoDB’s horizontal scaling capabilities allow the system to handle a large volume of resume data efficiently. As the number of users and resumes grows, MongoDB ensures that the system maintains performance by distributing data across multiple servers, providing both flexibility and high availability.To effectively analyze and process resumes, the system securely collects and organizes user-generated resume data into structured sections such as: Personal Details: Name, contact information, and location Education**:** Academic qualifications, degrees, institutions, and graduation dates,Work Experience: Details of ,previous employment, roles, and accomplishments,Skills**:** Technical and non-technical competencies.

This structured data is stored in MongoDB and serves as the foundation for subsequent processing phases. The categorization ensures that essential information is readily accessible for analysis while maintaining data integrity and security.

**4.4. MACHINE LEARNING MODEL**

**4.4.1. Data Collection and Storage**

**a) Classification Dataset**



**Source:** Kaggle

**Categories:** 24 distinct categories, including various professions and roles.

**Total Records:** 891 data points.

Columns:

**Resume:** Contains the textual data from resumes.

**Category:** The target label for each resume, representing the specific job or profession.

**b) Scoring Dataset**

**Custom Data:** Created manually to help score the resumes.

**Total Records:** 10,000 rows**.**

**Columns:**

**Experience:** The number of years of experience of the candidate.

**Skills:** The list of skills possessed by the candidate.

**Company Level:** The level/type of company where the candidate has worked.

**Score:** The final output column, scoring candidates on a scale of 0 to 100 based on the above attributes.

**4.4.2. Natural Language and Processing**

NLP techniques play a crucial role in extracting meaningful insights from resume data. Key steps in the preprocessing pipeline include:

**Tokenization**:

The text is split into individual words or tokens, creating a foundational structure for further analysis.

Example: "Software engineer with 5 years of experience" → ["Software", "engineer", "with", "5", "years", "of", "experience"].

**Stop-Word Removal**:

Commonly used but insignificant words (e.g., "is," "the," "and") are filtered out to focus on meaningful content.

This helps reduce noise and enhances the quality of extracted information.

**Lemmatization**:

Words are reduced to their base or root form, ensuring consistency in textual data.

Example: "running" → "run," "studies" → "study."

**TF-IDF For Text Vectorization**:

TF-IDF (Term Frequency-Inverse Document Frequency) is used to convert text into numerical vectors by measuring the importance of words. It combines term frequency (word occurrence in a document) with inverse document frequency (rarity across the dataset) to emphasize significant terms while reducing the weight of common words. In ResuPro, TF-IDF transforms resume text into vectors.

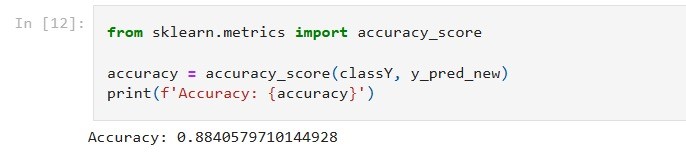
**4.4.3. Domain Classification**

The extracted features are passed to machine learning models to classify resumes into relevant job domains. The classification process involves:

**Naive Bayes Classification:** This probabilistic model is based on Bayes' Theorem and assumes that the features (words in the resumes) are independent. It calculates the probability of a resume belonging to each category and selects the category with the highest probability. Naive Bayes is particularly effective for text classification tasks due to its simplicity and efficiency, especially when dealing with large datasets.

**Random Forest Classification:** This ensemble method builds multiple decision trees and combines their results to improve accuracy and reduce overfitting. Each tree in the forest makes a prediction, and the final output is based on the majority vote of all the trees. Random Forest is known for its robustness and ability to handle complex data structures, making it a powerful tool for classification tasks like resume categorization.

**Accuracy:**



Both models were trained on the labeled dataset, allowing them to learn the patterns and relationships between the features of the resumes and their corresponding categories.

**4.4.4. Scoring Mechanism**

The Score Prediction Model evaluates and ranks candidates within a specific domain or job category by assigning a numerical score. Once a resume is classified into a domain, the model assesses three primary factors:

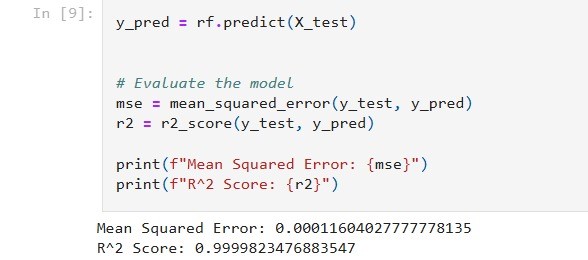
**Skills**: The relevance and presence of required skills.

**Experience:** Total years of professional experience.

**Company Level:** The type of company candidates has worked at, categorized into startup, mid-level, and enterprise.

The model uses a Random Forest Regressor, an ensemble learning algorithm that constructs multiple decision trees and aggregates their predictions for improved accuracy and reliability. It was trained on a manually curated dataset of 10,000 records, including fields for experience, skills, company level, and corresponding scores (ranging from 0 to 100).

**Accuracy:**



This scoring mechanism provides an objective evaluation of candidates’ qualifications, enabling effective ranking and comparison. By leveraging the model, recruiters gain insights into candidates' suitability for specific roles, streamlining decision-making and ensuring fair and informed selections.

**4.5. DEPLOYMENT**

ResuPro has been deployed on a highly scalable and secure server infrastructure designed to handle a large volume of users while ensuring high availability and optimal performance. The deployment architecture leverages cutting-edge tools and services for both frontend and backend, ensuring seamless integration and efficiency.

**4.5.1. Frontend Deployment on Netlify**

The frontend of ResuPro, built with React and styled using Tailwind CSS, has been deployed on Netlify. Netlify offers several benefits for modern web applications:

**Continuous Deployment**: Netlify supports automatic deployments from Git repositories (e.g., GitHub). Every change made to the repository triggers an automatic build and deployment pipeline, ensuring the website is always up to date without manual intervention.

**Global Content Delivery Network (CDN)**: Netlify leverages a global CDN, ensuring fast and reliable access to the frontend for users worldwide. The CDN caches content at the edge, reducing latency and improving load times for users, regardless of their geographical location.

**Serverless Functions**: For any lightweight backend functionalities required by the frontend (such as form submissions or user authentication), Netlify offers serverless functions that can scale automatically based on demand.

**High Availability**: By hosting on Netlify, the frontend benefits from its robust infrastructure, providing scalability and high availability without worrying about managing servers.

Netlify’s ease of use, coupled with its ability to automatically deploy changes and serve content quickly from a global network, ensures that ResuPro’s frontend remains responsive and reliable.

**4.5.2 ML MODEL DEPLOYMENT ON STREAMLIT**

The backend of ResuPro, which involves the machine learning (ML) model for resume classification, scoring, and ranking, is deployed on Streamlit. Streamlit provides an ideal environment for deploying data science and machine learning applications, offering several advantages:

**Rapid Deployment**: Streamlit allows for the quick deployment of Python-based applications with minimal setup. This is particularly useful for ResuPro’s ML models, which require Python for data processing and inference.

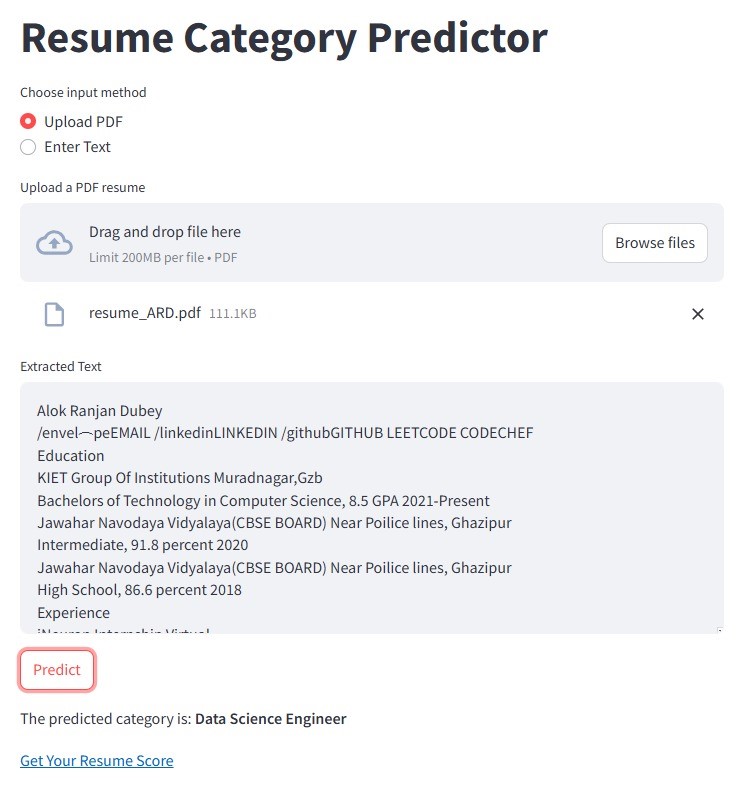
Streamlit's cloud service ensures that the backend can scale based on user demand, providing sufficient resources for handling large volumes of resume data and ML predictions.

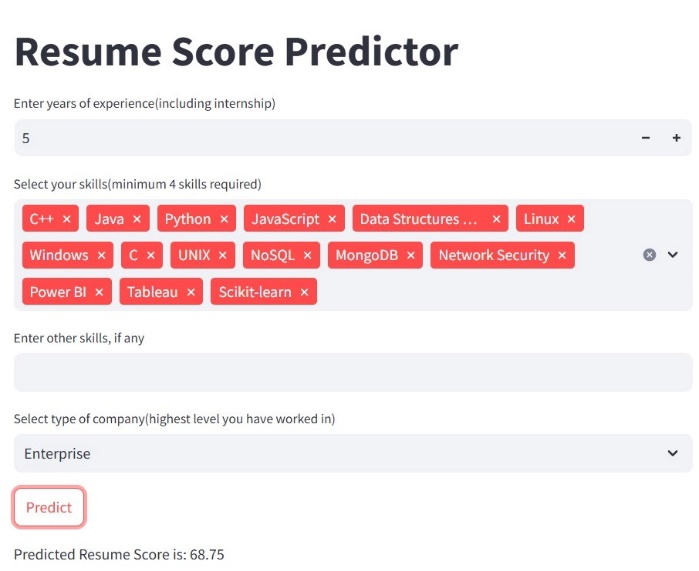
**Integration with Machine Learning Models:** Streamlit integrates seamlessly with Python libraries like Scikit-learn, TensorFlow, and PyTorch, making it ideal for serving ResuPro’s ML models (such as Random Forest and Naive Bayes) and running inference in real-time.

**Performance Monitoring:** Streamlit also supports continuous monitoring and logging, ensuring the backend remains performant and identifying any potential issues early.

By deploying the backend on Streamlit, ResuPro ensures that the system can efficiently handle the heavy computational requirements of the ML model, process resume data quickly, and provide real-time feedback to users.

# RESULT AND DISCUSSION





Initial testing of ResuPro has shown promising results in automating the resume screening process. The NLP module successfully extracts key information with an accuracy of over 90%, and the machine learning models achieve a classification accuracy of 88% across different job domains. The scoring mechanism effectively ranks resumes, providing useful insights for both candidates and recruiters. Feedback from pilot users indicates that the system significantly reduces the time required for initial resume screening and helps in identifying top-tier candidates more efficiently.

1. **CONCLUSION**

The ML-Based Resume Classifier (ResuPro) offers a significant improvement over traditional resume screening methods by automating the process using machine learning and NLP. The system enhances the recruitment process by providing a fast, scalable, and unbiased approach to resume classification and scoring. By integrating domain classification and a robust scoring mechanism, ResuPro helps recruiters make better-informed hiring decisions, ultimately improving the quality of talent acquisition.

This project represents a notable advancement in HR technology, addressing key challenges in talent acquisition. With its focus on efficiency and accuracy, ResuPro is poised to revolutionize the way organizations screen and evaluate candidates, making it a valuable tool for modern recruitment strategies.

1. **REFERENCES**
2. Surendiran B, Harsha Vardhan Chirumamilla, Maruprolu Naga, Raju Reddy. "Resume Classification Using ML Techniques." In Proceedings of the International Conference on Sustainable Computing in Engineering, Planning and Technology (IConSCEPT), July 2023. National Institute of Technology Puducherry. DOI: 10.1109/IConSCEPT57958.2023.10169907.
3. Rajath V , Riza Tanaz Fareed , Sharadadevi Kaganurmath, 2021, Resume Classification and Ranking using KNN and Cosine Similarity, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 10, Issue 08 ,August 2021
4. Jones, A., Smith, B., & Lee, C. (2020). Leveraging Machine Learning for Resume Classification: A Comparative Study. Journal of AI Applications in Recruitment, 8(2), 45-62
5. Obradović, M., Petrović, I., & Đurić, D. (2019). Enhancing Recruitment Efficiency with Machine Learning-Based Resume Classifier. Proceedings of the International Conference on Artificial Intelligence in HR, 123-135
6. Suhas. H. E and Manjunath. A E, “Differential Hiring using a Combination of NER and Word Embedding,” International Journal of Recent Technology and Engineering (IJRTE), vol. 9, no. 1. Blue Eyes Intelligence Engineering and Sciences Engineering and Sciences Publication - BEIESP, pp. 1344–1349, May 30, 2020.
7. Sujit. Amin, N. Jayakar, S. Sunny, P. Babu, M. Kiruthika and A. Gurjar, "Web Application for Screening Resume," 2019 International Conference on Nascent Technologies in Engineering (ICNTE), Navi Mumbai, India, 2019, pp. 1-7.
8. Pradeep Kumar Roy, Sarabjeet Singh Chowdhary, Rocky Bhatia, "A Machine Learning approach for automation of Resume Recommendation system", Procedia Computer Science, Volume 167, 2020, Pages 2318-2327, ISSN 1877-0509.
9. Bhushan Kinge, Shrinivas Mandhare, Pranali Chavan, S. M. Chaware, "Resume Screening using Machine Learning and NLP: A proposed system", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 8, Issue 2, pp.253-258, March-April-2022.
10. D. Jagan Mohan Reddy, S. Regella and S. R. Seelam, "Recruitment Prediction using Machine Learning," 2020 5th International Conference on Computing, Communication and Security (ICCCS), Patna, India, 2020, pp. 1-4
11. Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:181.04805.
12. Roe, R. A., Smith, K. M., & Smith, H. M. (2019). The Hidden Biases of Automated Resume Screening. Journal of Human Resource Management, 7(4), 12-29
13. Zhang, Y., et al. (202). The Future of Recruitment: A Machine Learning Approach. International Journal of Information Management, 50, 70-83.
14. Morley, C. S. Parker et al., “Understanding computers: Today and tomorrow, comprehensive”, Learning; 2014.
15. Prasanna Parasurama and João Sedoc, 2022: Gendered Information in Resumes and its Role in Algorithmic and Human Hiring Bias. Proceedings, 2022.
16. Tejaswini K, Umadevi V, Shashank M Kadiwal, Sanjay Revanna, Design and development of machine learning based resume ranking system, Global Transitions Proceedings,Volume 3, Issue 2, Pages 371-375, 2022
17. Smith, J., & Johnson, L. (2018). Machine Learning Approaches for Resume Screening: A Review. International Journal of Human Resources Management, 25(4), 567-580
18. Dr. Ajay K Shrivastava, Dr. Akash Rajak, Dr. Arun K Tripathi. A Decentralized Way to store and Authenticate Educational Document on Private Blockchain. 2019 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), Ghaziabad. ISBN:10.1109/ICICT46931.2019.8977633.
19. Anil Kumar Ahlawat, Gaurav Dubey, (2022-23). Aspect term extraction and optimized deep fuzzy clustering-based inverted indexing for document retrieval. Intelligent Decision Technologies. ISSN:1872-4981 (P), 1875-8843 (E)**.**