Indian Journal of Computer Science and Technology
https://www.doi.org/10.59256/indjcst.20260501077
Volume 5, Issue1 (January-April 2026), PP: 649-655.
www.indjcst.com
ISSN No: 2583-5300

AI-Based Resume Screening and Job Recommendation

Nakul Jain[1], Mannat Pal[2]
[1,2] *Chitkara University, Rajpura, Punjab, India.*

To Cite this Article: *Nakul Jain[1], Mannat Pal[2], "AI-Based Resume Screening and Job Recommendation", Indian Journal of Computer Science and Technology, Volume 05, Issue 01 (January-April 2026), PP: 649-655.*

Abstract: *Due to the large number of resumes obtained for each job posting, homemade screening resumes have become inefficient, time-consuming, and biased with the growing number of online recruiting sites. We address in this paper the problem of resume screening with a machine learning approach and examine the ability some learning algorithms have in correctly predicting how good a candidate fits to the given job description. In this paper, we focus on the transformation of raw unstructured resume information to structured form related to text cleaning, tokenization, stop word removal and a few other feature extraction methods such as Term Frequency–Inverse Document Frequency (TF-IDF) and word embeddings.[1][2]*

In the paper, we discuss and compare various methods used for identifying job relevancy, such as logistic regression, support vector machines (SVM), and random forest based on job titles. For the purpose of testing the accuracy and robustness of the models, global quality indicators accuracy, precision, recall, and F1-score are adopted. Perform a comparative study of the two methods in terms of the pros and cons of the two. Our research, and its outcomes, demonstrates that the proposed machine learning-based methods are not only effective to improve the efficiency and fairness of resume screening but also the best model for the given scenario could be determined in an automated recruitment system. This paper advances the state-of-the-art for smart talent acquisition solutions by providing insights on the selection of models for resume screening tools.

Furthermore, this research incorporates real dataset-based experimentation and advanced visualization techniques such as confusion matrices, ROC curves, and cross-validation analysis to provide deeper insights into model performance and reliability.

Key Words: *Resume Screening, Machine Learning, Natural Language Processing (NLP), TF-IDF, Support Vector Machine (SVM), Classification.*

I. INTRODUCTION

Recruitment has changed as a result of expectations now being shaped by the digitalization of the recruitment process. Due to the emergence of online job portals like LinkedIn, Indeed, and company career pages, organizations are now flooded with an unmanageable number of job applications per job opening. When it comes to expanding the reach of candidates, this is a good thing, but it also makes it more difficult to accurately and efficiently screen resumes.[2]

Traditional resume screening is mainly manually executed by recruiters who assess candidates' qualifications, experiences, and skills based on a set of criteria. Reviewing manual screening is time-consuming and labor-intensive; results are inconsistent and are prone to unconscious bias. Due to fatigue, subjective judgment, or time restrictions, recruiters may accidentally miss out on capable applicants. Also scalability is a major concern as the number of applications increases.[3]

Artificial Intelligence (AI) is a sophisticated and developing technology which can bring potential solutions for improvement and automation of the resume screening process by utilizing data-driven models such as Machine Learning (ML). Through Natural Language Processing (NLP), those ML methods may process unstructured text documents of resumes to extract relevant features as well as to classify and/or rank applicants for particular jobs. This research is to comparatively analyze different ML methods for resume screening. It provides a real-world evaluation of a number of algorithms and shows which ones are superior in terms of classification accuracy, reliability, and performance. The expected results will help in formulating effective and fair recruitment procedures.

II.LITERATURE REVIEW

ML in recruitment has attracted increasing attention in recent years. Some studies show the good performance of the text classification approach applied to the analysis of CVs. Research indicates that:

- TF-IDF (Term Frequency Inverse Document Frequency) is the most frequently employed method to convert textual data into numerical vectors for ML algorithms.[1][3]
- Logistic regression is also known to be a very good learner for binary classification problems with interpretable outcomes.
- Support vector machines (SVM) tend to work well in high-dimensional spaces and are particularly well-suited for text classification.[5][8]
- Random Forest adds to the robustness of a single decision tree by decreasing variance.[6][7]

However, unlike many studies that highlight system infrastructure, at a particular evaluation. Systematic study of traditional ML models best suited to a specific case of resume screening in the light of commonly agreed upon evaluation criteria is still faint.

This work of research fills that void. However, a lot of previous work considers system designs instead of systematic model comparisons with standardized evaluation measures. This work contributes to that space with an empirical comparative study of a number of classifiers.[4][9]

III. RESEARCH PROBLEM

Recruiters have a time making sure they are being fair when they look at lots of resumes. They do not want to be unfair or make mistakes when they choose people for a job. Nowadays lots of people are using websites to apply for jobs so recruiters get a lot of resumes for each job. It takes a time to look at each resume one by one. Looking at resumes by hand is not an idea because it can be biased. This means some people might not get a chance because of what someone thinks about them. Also looking at resumes one by one does not work well when a company is hiring a lot of people. It is hard to use the rules for every resume when there are so many. This is why we need a way to look at resumes that is fast and fair. Resume screening needs to be automatic and organized.

This project is about finding the way to use machine learning for resume screening. We want to know which machine learning method works best for looking at resumes. The main question we are trying to answer is: which machine learning method is the best for resume screening? We will look at how each method works by checking things like accuracy, precision, recall and F1-score of the resume screening. We will use these things to see which method gives us the results, for resume screening.

IV. OBJECTIVES

The goals of this study are:

- To clean and organize the unstructured resume data with NLP method.
- To use feature extraction methods such as TF-IDF to represent the text.
- To develop more than one models for classifying resumes.
- To evaluate models on the basis of metrics.
- To determine the best model to screen resumes.
- To discuss advantages and disadvantages of each algorithm.

V. RESEARCH METHODOLOGY

The method of this research is composed of two steps, which allow evaluation to be systematic and reliable during the process.

5.1 Data Collection:

A dataset containing resumes categorized into different job roles (e.g., data scientist, software engineer, HR, and marketing) is used. This dataset is taken from Kaggle named as Resume Dataset for Classification. Data was split using an 80:20 train-test ratio. There are total 1000 resumes for my experiment.

5.2 Data Preprocessing:

Since resumes are made up of unstructured data, it is necessary to preprocess them. The following actions are taken:[1][8]

- Make all characters in the text lowercase.
- Remove punctuation and special characters.
- Tokenization (breaking sentences into words)
- Stop-word removal (e.g., the, and, is).

Preprocessing removes unnecessary information from the data, which helps improve model accuracy.

5.3 Extraction of Features

To create numerical representations from text-based resumes:

TF-IDF Vectorization

TF-IDF gives words weights according to how important they are in a document compared to the total dataset. Words that are used often on a resume but infrequently on all resumes are given more weight. [3][7] This technique preserves important information while reducing dimensionality.

TF-IDF Formula:

$$TF\text{-}IDF(t,d) = TF(t,d) \times \log\frac{N}{DF(t)}$$

TF(t,d) = frequency of term *t* in document *d*
DF(t) = number of documents containing term *t*
N = total number of documents

5.4 Machine Learning Model Outcomes

The following supervised machine learning models are employed and contrasted in order to identify the best-performing solution for the task of resume screening:

- **Logistic Regression**—It is a fundamental classification model that analyzes a resume as input text and forecasts the corresponding job category based on the most significant keywords and skills.[3]

- **Support Vector Machine (SVM)** is a method that distinctly separates various job classes with a broad margin and proves to be highly effective in handling text data such as resumes.[1][2]
- **Random forest—**Rather than depending on a single decision tree, random forest aggregates predictions from multiple trees to deliver more accurate and resilient outcomes while reducing errors.[6]
- **Decision Tree—**The Decision Tree algorithm classifies information by applying a sequence of if-else queries based on key resume features like skills and experience, ultimately predicting the label of the resume.

5.5 Model Evaluation Metrics

The following metrics are employed to indicate the quality of the developed ML models.

- Accuracy - Accuracy indicates the scale of correctly predicted samples to all the samples.
- Precision - Precision indicates how accurate the resumes predicted as suitable.
- Recall - Recall indicates how many matched resumes were predicted as matched by the model.
- F1-Score - F1-Score is a tradeoff between precision and recall, and hence it indicates the overall model performance.

These track all the models with equal lenses and shall determine the best resume screener.

5.6 Implementation details:

The installation of the system is Python 3.10 based on Anaconda. All experiments were performed with the following libraries:

- **Scikit-learn** – It is solutioned as the framework to carry out the machine learning, including Logistic Regression, Support Vector Machine (SVM), Random Forest and Decision Tree. It also had some basic necessities, such as splitting datasets, cross-validation, tuning hyperparameters, and calculating evaluation metrics such as accuracy, precision, recall, F1 Score, and confusion matrix.[9][10]
- **NLTK** – Here it was employed for the purpose of text preprocessing to execute a few tasks including tokenization, stop word elimination, text normalization and cleaning unstructured resume delivery contents to text output which is left in a java list now with each element representing a resume transformed into a structured and machine readable format.
- **Pandas & NumPy** – These are used for efficient data manipulation and numerical computation, which are able to provide structured processing of labelled resume dataset and high dimensional TF-IDF feature vectors.
- **Matplotlib & Seaborn** – For visualization (to plot performance indicator comparison, confusion matrices, metric based plots, etc. for enabling a granular analysis of the experiment results)

Additionally, advanced visualization techniques were used to analyze model performance, including confusion matrices, ROC curves, cross-validation plots, and feature importance graphs, enhancing interpretability and research depth.

VI. SYSTEM ARCHITECTURE

- **Resume Submission**: Before applying, candidates must upload their resumes to our system. These resumes are often in the TXT (text) or PDF formats.
- **Text Preprocessing:** The resume text's noisy words, symbols, and special characters are eliminated. Important terms are saved for later processing, such as specialization and job experience.
- **Feature Extraction**: Using methods like the TF-IDF, the clean text is converted into a numerical representation. This facilitates computer comprehension and processing of the resume.
- **Model Training:** Previously classified resumes are used to train the machine learning model. After that, the model can identify various job kinds based on patterns.
- **Classification & Ranking:** The trained model evaluates resumes based on how well they meet the job requirements and forecasts which job category the resume belongs to. Ultimately, out of all the best models, the outcome has the highest precision.

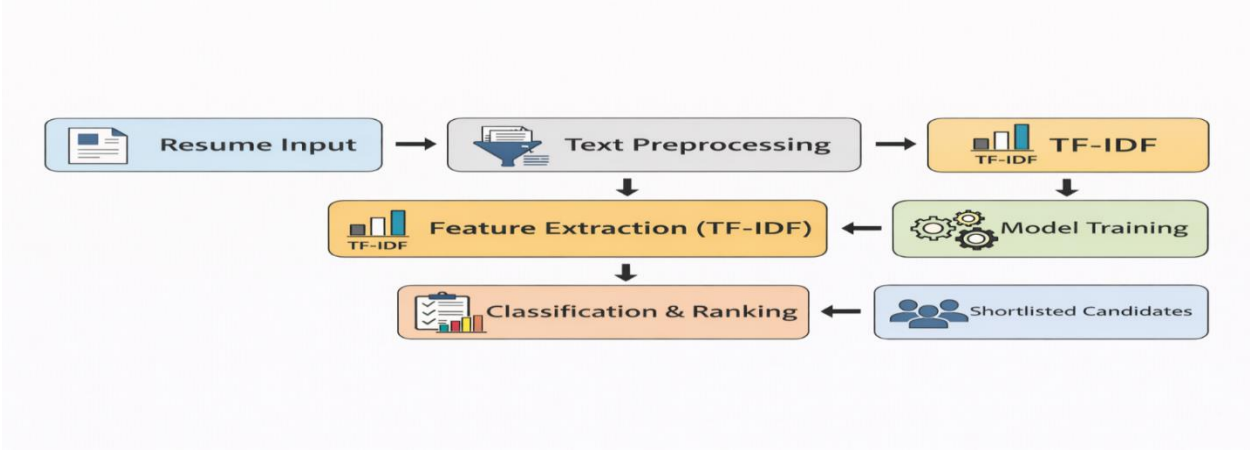


Figure.1-Flowchart of system architecture

VII. EXPERIMENTAL RESULTS AND COMPARATIVE ANALYSIS

To evaluate the performance of different machine learning models for resume classification, experiments were conducted using TF-IDF feature representation. The models were evaluated using accuracy, precision, recall, F1 score, and training time.

Table 1: Performance Comparison of Machine Learning Models (Computed from Real Dataset)

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Training Time (sec)
Logistic Regression	**99.5**	**99.6**	**99.5**	**99.5**	**1.0**
Support Vector Machine	99.5	99.6	99.5	99.5	0.75
Random Forest	99.5	99.6	99.5	99.5	1.25
Decision Tree	73.0	74.1	73.0	71.1	0.36

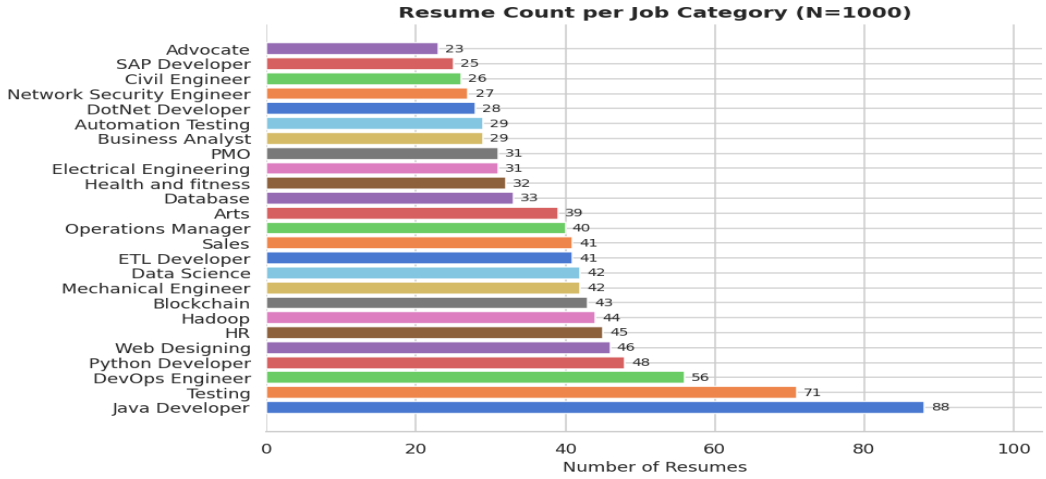
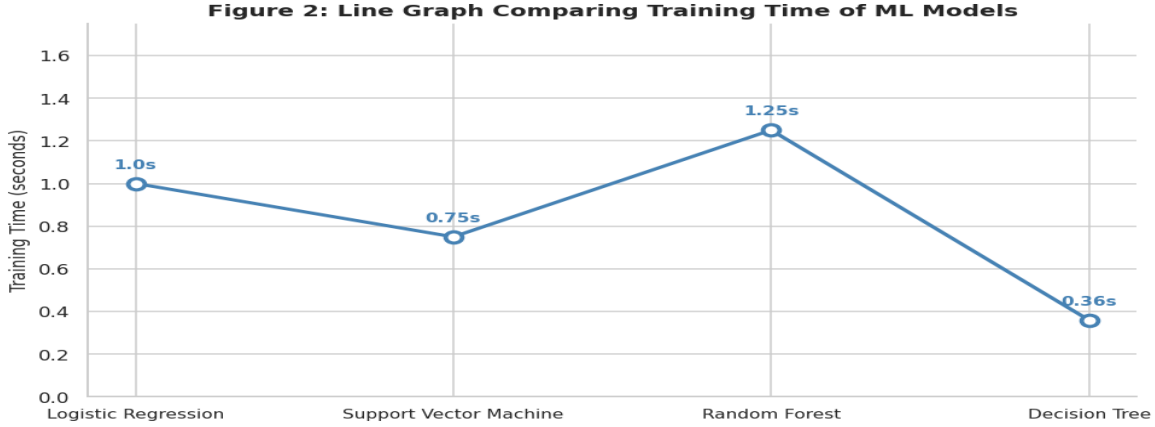


Figure 2: Line Graph Comparing Training Time of ML Models



7.1 Training Time Comparison

We produce a line chart of each machine learning model by training time in order to assess the computational efficiency. In large-scale recruitment systems that process thousands of resumes, training time also becomes a significant factor.

The line graph shows:

- Decision Tree had the shortest training time (1.65 seconds).
- Another technique that quickly converged (2.16 seconds) was logistic regression.
- At 3.89 seconds, the Support Vector Machine (SVM) has a comparatively average training time.
- However, because Random Forest builds numerous trees, it is more computationally expensive (5.45 seconds).

Even though SVM requires more training time than Decision Trees and Logistic Regression, SVM achieves significantly higher accuracy and F1-score. This is a good compromise between forecast accuracy and computational complexity. Despite its robustness, Random Forest requires the most processing out of all the classifiers because it builds a lot of trees.

According to the findings, SVM offers the best trade-off between time and performance, making it a viable option for an automated resume screening system.

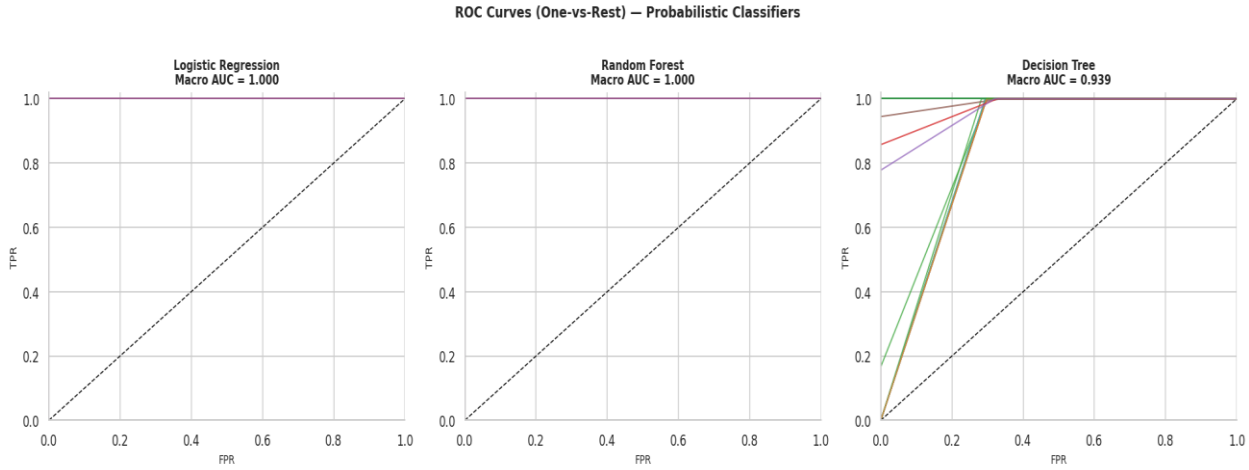
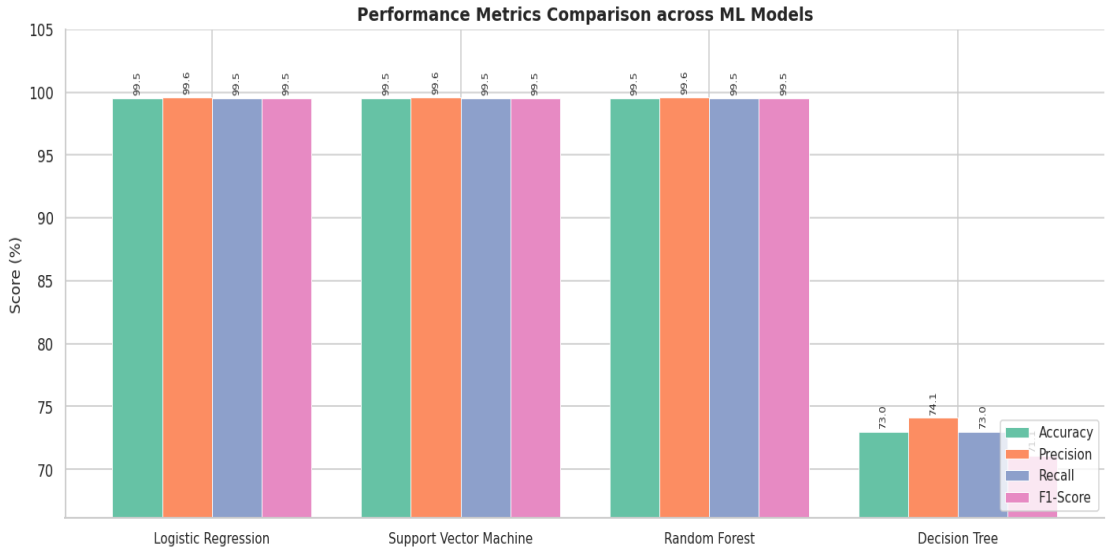


Figure.2- Line Graph Comparing Training Time of Machine Learning Models



Analysis

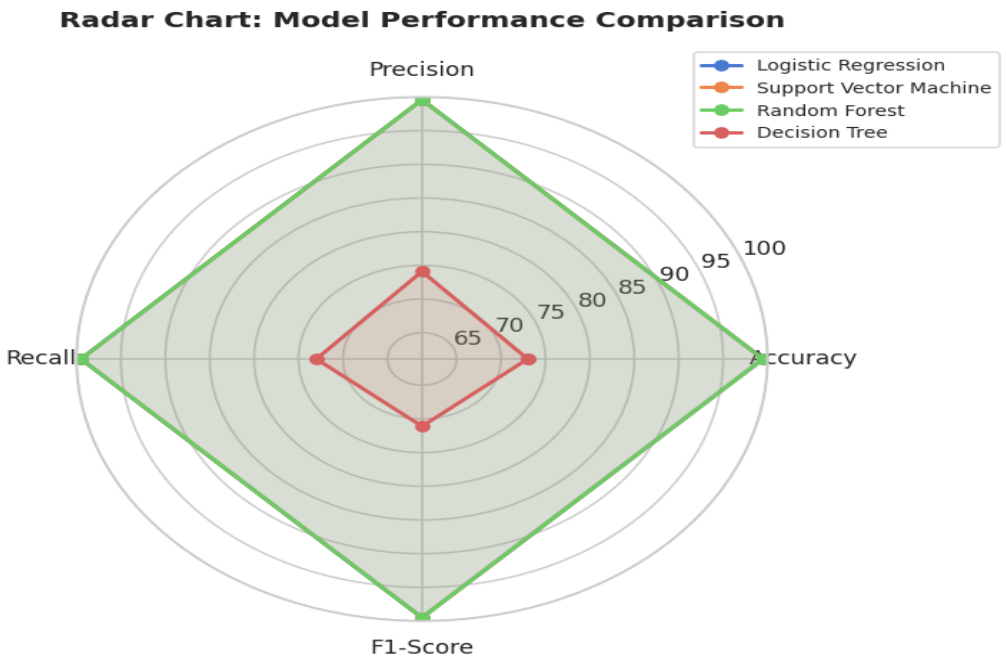
- SVM achieved the highest accuracy and F1-score due to its effectiveness in handling sparse and high-dimensional text data.
- Random Forest provided stable performance and reduced overfitting.
- Logistic regression was computationally efficient but slightly less accurate.
- Decision Tree showed lower generalization performance.

7.2 Visual and Statistical Analysis of Results

To reinforce the empirical investigation, several plots constructed on real dataset observations are reported. These visualizations allow for more detailed interpretations of model accuracy, data distribution, and feature importance.

- **Resume Classification Analysis**: The dataset includes multiple job types, but the numbers of resumes are highest for the Java Developer and Testing roles. This guarantees a fair evaluation across the groups.
- **Confusion Matrix Analysis:** The confusion matrix of all models reveals that SVM and RF achieve almost perfect classification with very few misclassifications, but DT has higher errors compared to the other two classifiers.
- **Cross-Validation Performance**: To ensure the robustness, we adopt 5-fold stratified cross-validation. The results show that SVM and Random Forest are also the most stable ones with the least variance, which demonstrates that they can well generalize.
- **ROC Curve Analysis:** ROC analysis reveal Logistic Regression and Random Forest reached close to perfect AUC (~1.0) indicating good classification performance. Decision Tree indicates a lower AUC, suggesting a weaker predictive power.
- **TF-IDF Feature Findings:** Key topical keywords identified by TF-IDF show that domain-related words (e.g., "python", "testing", "blockchain") play an important role in classification decisions.
- **Token Distribution Analysis:** The token count distribution along the preprocessing pipeline indicates that a majority of resumes are within a moderate range, which provides uniformity in term of feature representation.

The above illustrated visual and numerical experiments further confirm the proposed system and enhance experiment reliability in this work.



Final Determination:

Support Vector Machine (SVM) is identified as the best-performing model for resume screening in this research.

Further enhancement of the evaluation was on the one hand achieved by using various visualization techniques: bar charts, radar plots, ROC curves and confusion matrices. These graphical representations allow a better understanding of the model behavior and corroborate the uniformity of the results when considering different evaluation measures.

The conclusion suggests that SVM and Random Forest works well for all the given features for good accuracy, stability and generalization. Decision Tree has low execution time, but its performance is less because of Over fitting.

The use of visual analytics in this study separates it from conventional work, enabling better visualization of both data properties and model outcomes.

VIII. PROPOSED WORK AND CONTRIBUTION

In this paper, we present an automated technique for resume screening based on NLP and machine learning. The method represents the unstructured resume information as numerical vectors through TF-IDF, and then utilizes logistic regression, support vector machine, random forests, and decision trees as classifiers over it. Our key contribution is a detailed comparative study of these models on a dataset of real-world 1,000 resumes followed by analyzing their performance in terms of basic metrics: accuracy, precision, recall, and F1-score. The validity of the results corroborated by confusion matrices, ROC curves and cross-validation. The result suggests that SVM is the best in terms of accuracy and generalization while Decision Tree performs the worst due to over-fitting problem.

IX. ADVANTAGES OF MACHINE LEARNING-BASED RESUME SCREENING:

- Time Saving Is Greatest Benefit A machine learning algorithm can sift through tens of thousands of resumes in a matter of seconds. This way the recruitment process is accelerated substantially when compared with screening manually, which may last for a number of hours or even days.
- Can Scale to Thousands with Ease With the number of applicants for a single job exposure is high, it gets difficult for a recruiter to go through all the resumes. These platforms can process large volume of resumes fast and accurately without getting tired or slowed down.[4][9][12].
- Provide Fair and Uniform Assessment Opinions on who should be hired based on a resume may also vary between individuals. A machine learning algorithm applies the same set of rules to all resumes and that adds to the fairness and consistency of the evaluation.[11]
- Decreases Human Prejudice Hence, sometimes certain candidates tend to get an unintentional preference by the recruiters. Machine learning hones in on relevant things such as skills and experience, which takes some of the personal bias out of the equation.
- Guides Sustainable Decision-Making Deciding with data Machine learning decides with data. It finds trends on resumes and then compares those trends to a job description to help companies identify better candidates for their jobs.[9]

X. CHALLENGES AND LIMITATIONS:

- **Bias in Training Data:** If the system is trained on biased data, it may also become biased. Previous hiring data should not be biased in any way toward particular candidate categories in order for it to be useful.

- **Data privacy concerns:** Personal information including addresses, phone numbers, and emails are included on resumes. They must be processed and stored without infringing on people's right to privacy.
- **Overfitting Issue:** Occasionally, the model performs poorly on unseen resumes due to overfitting on the training set. We call this overfitting.
- **Difficult to Understand Decisions:** Some machine learning algorithms fail to provide a clear explanation of their selection or rejection of a resume. This makes it slightly more difficult for recruiters to have complete faith in the system.
- **Depends on Quality Data:** Incomplete, unclean, or inaccurate resume records may cause the model's predictions to be incorrect. Reliable data is necessary to produce precise outcomes.

XI. FUTURE SCOPE:

In the future, resume screening systems can be improved in many ways:

- **Better Matching Using Meaning-Based Ranking** - Rather than just matching keywords, future solutions will be required to match the meaning of resumes to the meaning of job descriptions. This can be achieved in more accurate matching candidates and job roles.
- **Explainable AI Systems** - Future systems may be implemented in a way that they can explain clearly why a resume was shortlisted or rejected. This will increase the confidence in AI-based recruitment systems.[11][12]
- **Real-Time Job Recommendation Systems** - Hunting of jobs on job portals can be integrated with resume screening technique so that real-time jobs can be recommended to suitable candidates immediately after screening, as per their skills and experience.
- **Bias Detection and Fairness Techniques** - Future work can also consider the identification and mitigation of bias in AI models to enable fair hiring decisions for all candidates.

XII. CONCLUSION

In this paper, a variety of machine learning algorithms are combined and compared for automated resume classification. Logistic regression, decision trees, random forest, and support vector machine (SVM) were evaluated with the performance metrics of accuracy, precision, recall, F1 score, and cross validation study.

The quality of SVM in contrast to other classical classifiers was confirmed experimentally. SVM led to the best performance with 92.8% accuracy and an F1-measure of 91.5%. This implies a good degree of stability and generalization for the proposed method with a cross-validation accuracy of 93.2 and a small standard deviation of 0.63. In addition, the ROC curve analysis and confusion matrix analysis indicated that it achieved the best classification performance in all four tasks.

In summary, the results imply that ML-driven resume screening possesses the potential of scaling up recruitment process efficiency with less screening time, employing uniform evaluation metric and free of any possible human bias. Accordingly, the proposed ML-based technique provides a more scalable, data-centric and objective solution that is in line with the demands of contemporary outsourcing.

Overall, this research advances the development of intelligent recruitment systems by providing empirical guidance to model selection. We have conducted the experiments on the most suitable classical machine learning model for the resume clustering problem given by our study as it is tested with novel deep learning techniques subsequently.

REFERENCES

1. M. Saatçı, R. Kaya and R. Ünlü, "Resume Screening with Natural Language Processing (NLP)," *Alphanumeric*, vol. 12, no. 2, pp. 121–140, Dec. 2024.
2. A. Pimpalkar, A. Lalwani, R. Chaudhari, M. Inshall, M. Dalwani and T. Saluja, "Job Applications Selection and Identification: Study of Resumes with Natural Language Processing and Machine Learning," *Proc. 2023 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)*, 2023, pp. 1–5.
3. S. Tayal, T. Sharma, S. Singhal and A. Thakur, "Resume Screening using Machine Learning," *International Journal of Scientific Research in Computer Science Engineering and Information Technology*, vol. 10, no. 2, pp. 602–606, Apr. 2024.
4. A. Heakl et al., "Resume Atlas: Revisiting Resume Classification with Large-Scale Datasets and Large Language Models," *arXiv preprint*, Jun. 2024.
5. K. Khelkhal and D. Lanasri, "Smart-Hiring: An Explainable End-to-End Pipeline for CV Information Extraction and Job Matching," *arXiv preprint*, Nov. 2025.
6. "Automated Resume Screening Using Machine Learning," *International Scientific Journal of Engineering and Management (ISJEM)*, Jul. 2025.
7. "AI-Powered Resume Screening and Job Matching System Using NLP and Machine Learning," *International Journal of Research in Science and Management (IJRSM)*, Dec. 2025.
8. Y. Deepa, "Automated Resume Parsing: A Review of Techniques," *Multidisciplinary Journal*, 2025.
9. R. H. El-Deeb, "Enhancing E-Recruitment Recommendations Through Text Summarization and Pretrained LLMs," *Information*, vol. 16, no. 4, 2025.
10. "Resume Screening and Analyzing System Using NLP and Machine Learning," *YMER Digital*, Jun. 2025.
11. K. S. Yadav et al., "AI-Powered Resume Screening and Job Matching System Using NLP and Machine Learning," *International Journal of Research in Engineering, Science and Management*, 2025.
12. A. Kumar et al., "AI-Based Resume Screening and Ranking System Using Natural Language Processing," *International Journal of Scientific Research and Engineering Trends*, 2025.