Adaptive Integration of Optuna Optimization and Stacking Ensemble Learning for Automated Work Competency Classification

Mutiana Pratiwi[1,*], Sarjon Defit[2], Muhammad Tajuddin[3]

[1]*Information System, Universitas Putra Indonesia YPTK Padang, Lubuk Begalung Main Street, Padang, 25221, Indonesia*

[2]*Information Technology, Universitas Putra Indonesia YPTK Padang, Lubuk Begalung Main Street, Padang, 25221, Indonesia*

[3]*Information Technology Department, Universitas Bumi Gora, Ismail Marzuki, Cilinaya Street, Cakranegara, Mataram, 83127, Indonesia*

(Received: November 15, 2025; Revised: January 15, 2026; Accepted: March 22, 2026; Available online: April 4, 2026)

Abstract

Artificial intelligence and machine learning are increasingly used to automate analytical and decision processes, including the evaluation of human competencies. However, traditional models often face challenges in accuracy and generalization when applied to linguistic data from interviews. This study aims to develop a model that integrates Optuna optimization and stacking ensemble learning to enhance the accuracy and robustness of competency classification. Interview transcript data were processed using natural language processing techniques such as cleaning, tokenization, case folding, stopword removal, and stemming to ensure textual consistency. The text was then transformed into numerical representations using term frequency inverse document frequency weighting. To handle class imbalance, the synthetic minority oversampling technique was employed. Optuna was applied to optimize the hyperparameters of base models, including support vector classifier, naive Bayes, random forest, gradient boosting, and XGBoost. These optimized models were combined through a stacking ensemble to form the final classifier. The dataset consists of 1 audio interview file in .txt format, containing 3,945 lines with 22,321 characters. To address class imbalance, SMOTE was applied, expanding the dataset to 850 samples. The model evaluation employed a 80:20 train–test split strategy combined with 10-fold cross-validation. The proposed model achieved an accuracy of 94 percent and a precision of 95 percent with macro and weighted F1 scores of 0.94. The results demonstrate stable and balanced performance across all competency categories, including analytical thinking, initiating action, problem solving, and work standards. Comparative analysis with previous studies in sentiment analysis, medical diagnosis, and financial forecasting confirmed that the integration of Optuna and stacking produces more robust and generalizable outcomes. The integration of Optuna optimization and stacking ensemble learning effectively improves classification performance while maintaining interpretability. The model demonstrates strong potential for automated competency evaluation in recruitment and human resource analytics. This framework can be extended to other linguistic datasets to support transparent and data-driven decision-making in artificial intelligence applications.

Keywords: Artificial Intelligence, Machine Learning, Natural Language Processing, Optuna, Stacking Ensemble

1. Introduction

The rapid advancement of Artificial Intelligence (AI) and Machine Learning (ML) has significantly transformed Human Resource Management, particularly in automating competency assessment [1]. Conventionally, evaluating competencies from interview transcripts is a manual, labor-intensive process that is often prone to subjectivity and inconsistency [2]. To address these limitations, ML approaches offer a robust solution by enabling the automated extraction of patterns from unstructured text data [3]. However, applying these models to interview transcripts presents specific challenges, such as handling high-dimensional linguistic features and ensuring classification accuracy in the absence of structured numerical inputs [4]. While Natural Language Processing (NLP) has been successfully applied to various domains such as sentiment analysis and spam detection, its application to competency classification from interview transcripts remains underexplored [5]. Unlike short structured texts, interview transcripts are characterized by high dimensionality, semantic sparsity, and irregular sentence structures [6]. Previous studies relying on single-model classifiers often struggle to capture the complex dependencies within such verbose data, leading to suboptimal generalization [7]. This limitation highlights the necessity for a more robust approach, such as ensemble learning, which can synthesize diverse predictive patterns to improve classification performance on unstructured long-form text

***Corresponding author: Mutiana Pratiwi (mutiana_pratiwi@upiyptk.ac.id)**
DOI: https://doi.org/10.47738/jads.v7i2.1228

[8]. Essentially, TC helps transform unstructured textual data into structured, interpretable information by assigning appropriate labels to each text segment [9]. This capability is vital for businesses seeking to derive strategic insights from textual data, support decision-making, and refine product and service strategies [10].

Natural Language Processing (NLP) is an interdisciplinary domain emerging from linguistics and computer science, primarily within AI [11]. NLP focuses on enabling computers to understand, interpret, and generate human language in a valuable way [12] . Rather than merely converting speech to text, NLP seeks to comprehend meaning and context to generate appropriate responses. Moreover, NLP supports advanced techniques such as Information Extraction (IE), which converts unstructured textual content into structured, analyzable information [13]. Information Extraction (IE) involves identifying entities, relationships, and events from unstructured texts, thereby producing structured datasets for downstream applications such as knowledge bases or decision-support [14]. Prior research has demonstrated the wide applicability of ML and NLP in diverse TC scenarios—such as AI-driven recruitment, personality analysis, and spam-email detection—laying a strong foundation for developing more efficient computational methods for automated competency assessment [15].

The integration of ML and NLP has been extensively applied in assessment processes, particularly in evaluating human competencies [16]. In corporate recruitment, ML models can predict candidate competencies by analyzing textual or verbal responses, thereby reducing human bias and improving efficiency. Such systems help minimize discrimination, save time and cost, and enhance candidate quality through data-driven evaluation [17]. Several prior studies have provided methodological and empirical foundations for this work. Proposed a hybrid feature-selection approach combining the Chi-Square filter and the Artificial Bee Colony (ABC) algorithm for Arabic text classification. Their approach reduced dimensionality by up to 94% and significantly improved weighted F1-scores using Naïve Bayes and SVM classifiers [18]. Then developed AraBig5, an ML framework employing logistic regression and SVM to predict the Big Five personality traits from Arabic tweets, demonstrating ML's potential for psychological inference [19]. Conducted a comparative study on fairness in AI-driven hiring systems, highlighting the role of models such as Random Forest, Decision Tree, KNN, AdaBoost, and XGBoost while emphasizing the ethical implications of algorithmic bias [20]. Similarly, applied SVM, KNN, and Multinomial Naïve Bayes (MNB) for social-media-based personality assessment using the MBTI and Big Five frameworks, providing recruiters with automated insights into candidate compatibility [21]. Developed Smart-Hire, a system combining KNN, CNN, and logistic regression for personality prediction and skill evaluation, enabling real-time personality identification and self-assessment [22].

In the health domain, introduced an ML-based framework for classifying mental health and emotions from English text, emphasizing its potential for early depression detection through emotion recognition [23]. Examined filtering-based feature selection for email-spam detection using the KNN classifier, revealing that modified PMI-based methods achieved 98.06% accuracy and 96.67% F-measure [24]. Compared word-embedding techniques (Word2Vec, GloVe, FastText) for Italian news categorization, where SVC achieved 93% F1-score on RCV2 dataset, showing the effectiveness of pre-trained embeddings [25]. Proposed the De-Redundancy Relative Discrimination Criterion (DRDC) for feature selection, combining Mutual Information with RDC to maximize relevance while minimizing redundancy, yielding superior Micro- and Macro-F1 scores on R8 and 20 Newsgroup datasets [26]. The proposed method aims not only to improve prediction accuracy but also to ensure model robustness when dealing with high-dimensional data from interview transcripts, addressing the limitations of single-model approaches.

2. Literature Review

Recent developments in Machine Learning (ML) and Natural Language Processing (NLP) have significantly advanced text classification, personality prediction, and recruitment analytics. Compared Random Forest and XGBoost for personality prediction from resumes, finding that Random Forest achieved higher accuracy (90%) than XGBoost (86%)[27]. Similarly, Proposed the Relevant-Based Feature Ranking (RBFR) method, improving classification accuracy by 25.43% across multiple ML models [28].

In AI-based hiring systems, applied ML models such as SVM, Naïve Bayes, and Decision Tree to analyze social media data for candidate assessment, demonstrating effective automation in personality evaluation [20][21]. Combined ML and NLP for emotion and mental health detection from text, while achieving superior accuracy by integrating BERT and XGBoost for medical text classification [23] [29]. Other studies extended NLP applications across diverse domains: Proposed a weighted ensemble classifier for malicious link detection; [30]

Sun and Luo [31] utilized Random Forest and Logistic Regression to analyze English writing text features. While

effective for structured essays, their approach relies heavily on manual feature engineering. In the medical domain, Rabbi et al. [32] and Ali et al. [33] demonstrated the superiority of stacking ensemble classifiers for risk prediction. However, these studies primarily focused on structured numerical datasets and fixed hyperparameters. Drawing from these findings, our current study adapts the stacking architecture for unstructured interview data and integrates Optuna to automate the hyperparameter tuning process, addressing the limitations of manual configuration found in prior works. However, a critical gap remains in the current literature regarding the application of ensemble-optimization frameworks to unstructured long-form text. While recent studies have successfully applied stacking and parameter optimization in domains such as medical diagnostics, and short-text sentiment analysis [11], [35], these approaches often rely on structured numerical data or limited linguistic contexts. They rarely address the specific challenges of interview transcripts, which are characterized by high dimensionality, semantic sparsity, and class imbalance. This study bridges this methodological gap by explicitly tailoring an Optuna-optimized stacking architecture to process TF-IDF weighted features from interview data, ensuring that the model captures complex linguistic patterns that traditional single-model optimization strategies fail to generalize.

To provide a clear overview of the research landscape and highlight the methodological contributions of this study, table 1 presents a structured comparison between related works and our proposed approach. As observed, while ensemble learning is widely used, the integration of automated hyperparameter optimization (Optuna) specifically for high-dimensional interview text data remains a distinct gap addressed by this research.

Table 1. Comparison of the proposed method with existing related studies

Author (Ref)	Dataset / Domain	Feature Representation	Method / Classifier	Optimization Strategy	Gap / Limitation
Sun & Luo [31]	English Writing Text (Structured Essays)	N-Gram / Statistical Features	Random Forest & Logistic Regression	Grid Search (Standard)	Focuses on structured text; limited handling of semantic sparsity.
Rabbi et al. [32]	Cardiovascular Data (Numerical/Medical)	Numerical Features	Stacking Classifier	Manual Tuning / Grid Search	Applied to structured numerical data, not applicable to unstructured text.
Ali et al. [33]	Diabetes Datasets (Numerical/Health)	Numerical Features	Stacking (RF + Meta Classifier)	Manual / Grid Search	Lacks automated hyperparameter tuning; prone to local optima.
Proposed Method	Unstructured Interview Transcripts (Work Competency)		TF-IDF (High-dimensional Text)	Stacking Ensemble (RF, SVM, NB, KNN)	Optuna (Automated TPE-based Optimization)
Author (Ref)	Dataset / Domain	Feature Representation	Method / Classifier	Optimization Strategy	Gap / Limitation

3. Methodology

The proposed research adopts an integrated Natural Language Processing and Machine Learning framework to classify and predict individual work competencies based on interview transcript data [36]. The entire process begins with data acquisition, where recorded voice interviews are transcribed using an automated speech-to-text system [37]. Each transcript is segmented carefully to preserve the semantic structure between interviewer questions and participant responses, forming the primary dataset for further processing [38]. The overall research process that examines the improvement of work competency classification through optuna-optimized ensemble stacking learning is illustrated in figure 1.

The dataset used in this study consists of 1 audio interview file in .txt format, containing 3,945 lines with 22,321 characters unstructured interview transcripts collected from job applicants in the academic sector. The interviews were conducted in Indonesian language, with an average duration of 30 minutes per session. To ensure the validity of the competency labels Target Variable, the transcripts were manually annotated by 3 subject matter experts, specifically senior HR managers or certified psychologists, each with a minimum of 5 years of experience in recruitment. The dataset consists of Due to the imbalanced distribution of competency classes, we applied SMOTE to the training set, resulting in a balanced total of [Y] samples. For the evaluation, the dataset was partitioned using a stratified [80:20] split, allocating [80%] for training and [20%] for testing.

While SMOTE is effective for balancing class distributions, its application to high-dimensional sparse data (such as TF-IDF vectors) presents specific challenges. The interpolation of minority samples in this space may generate synthetic feature vectors that do not correspond to linguistically valid sentences, potentially introducing synthetic sample distortion. However, this approach was prioritized over undersampling to avoid the loss of critical information in an already limited dataset. To mitigate the risk of noise introduced by synthetic samples, this study employs a

 ISSN 2723-6471

Stacking Ensemble framework. By aggregating predictions from diverse base learners (e.g., Random Forest and SVM), the ensemble model reduces the variance and sensitivity to individual noisy samples, thereby maintaining generalization performance despite the potential imperfections in the synthetically generated minority instances.

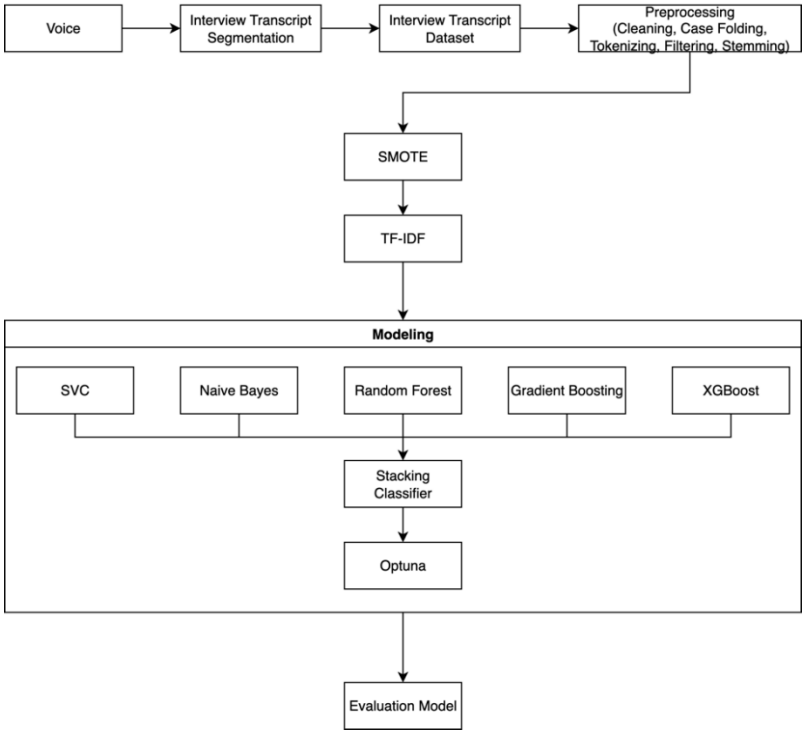

Figure 1. Methodology Flow.

Furthermore, to validate the model's reliability and mitigate overfitting, we implemented a [10-fold] cross-validation procedure during the hyperparameter tuning phase. The hyperparameter optimization is performed using the Tree-structured Parzen Estimator (TPE), which models the probability of hyperparameter configurations given their performance scores. The TPE algorithm models $p(x|y)$ using two non-parametric densities, as defined in Equation (1):

$$p(x \mid y) = \begin{cases} l(x), & \text{if } y < y^* \\ g(x), & \text{if } y \geq y^* \end{cases} \quad (1)$$

Where: x represents the hyperparameter configuration, y represents the value of the objective function (validation loss) to be minimized, y^* is the threshold value, typically set as the quantile γ of the observed objective values, $l(x)$ is the density formed by observations where the loss was lower (better) than $y*$. $g(x)$ is the density formed by observations where the loss was higher (worse) than $y*$.The algorithm selects the next hyperparameter set by maximizing the Expected Improvement (EI), which is proportional to the ratio $\frac{l(x)}{g(x)}$.

The text data are then pre-processed to ensure consistency and remove linguistic noise. The procedure involves cleaning to eliminate unnecessary symbols, case folding to normalize capitalization, tokenization to separate sentences into words, stopword removal to exclude non-informative terms, and stemming to reduce words to their root form. This stage produces a uniform dataset that improves feature extraction and model interpretability. Because interview datasets often exhibit class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) is applied to generate synthetic samples for the minority categories, ensuring that the model learns from balanced and representative data distributions. Following preprocessing, Term Frequency–Inverse Document Frequency (TF-IDF) is employed to convert textual data into numerical features. TF-IDF assigns weights to each term based on its relative importance across the corpus, enabling the model to focus on discriminative patterns that contribute meaningfully to competency identification. To enhance the generalization and stability of the classification model, this study integrates Optuna-based hyperparameter optimization with a Stacking Ensemble Learning strategy. Optuna, a Bayesian optimization framework, automatically searches for the optimal set of hyperparameters that minimizes the validation loss of each base learner. Formally, the optimization process is expressed as

 ISSN 2723-6471

$$\theta_i^* = arg \min_{\theta_i} \mathcal{L}(f_i(V; \theta_i)) \quad (2)$$

where f_i represents the i^{th} classifier trained on the TF-IDF feature matrix $V; \theta$ denotes its hyperparameter set, and L is the loss function to be minimized. This adaptive procedure enables each classifier to achieve its best performance without manual tuning. Once the optimal parameters are obtained, the optimized base models, such as Support Vector Classifier, Naïve Bayes, Random Forest, Gradient Boosting, and XGBoost, are combined through a Stacking Ensemble structure. The meta-learner aggregates the prediction outputs of all optimized base classifiers to generate the final decision, defined as

$$\hat{y} = g(f_1^*(V), f_2^*(V), \dots, f_m^*(V)) \quad (3)$$

where $f_i^*(V)$is the optimized output of each base model and $g(\cdot)$ represents the meta-classifier that learns higher-level representations from the ensemble. This hierarchical fusion allows the model to capture both linear and nonlinear decision boundaries, thereby improving robustness and predictive accuracy.

Model performance was evaluated using several quantitative metrics, including Accuracy, Precision, Recall, and F1-Score. These indicators provide a comprehensive understanding of both the overall and per-class performance of the system. The evaluation process ensures that the model can not only achieve high accuracy but also maintain balance across all performance dimensions, reflecting its robustness and reliability in classifying various competency levels.

The methodological novelty of this framework lies in its adaptive integration of Optuna and Stacking Ensemble Learning, which transforms hyperparameter tuning and model aggregation into a unified optimization process. Unlike previous works that apply these techniques separately, this study combines them systematically to create a more dynamic, explainable, and data-driven classification model. The result is an approach that achieves higher accuracy, improved generalization across linguistic variations, and reduced manual intervention contributing to a replicable and scalable solution for automated competency evaluation.

4. Results and Discussion

This section presents the outcomes derived from the implementation of the methodology described in the previous section, along with a systematic discussion and analysis of the research findings. The purpose of this section is to evaluate how effectively the developed model performs its classification function in accordance with the research objectives. Furthermore, the discussion connects the experimental results with relevant theories and previous studies to strengthen the scientific validity and relevance of the proposed approach. The presentation of results is organized to demonstrate the model's performance evaluation process, the effectiveness of the applied algorithms, and the methodological contribution toward improving the accuracy of competency classification based on interview data. The discussion analytically interprets the emerging patterns from the experimental results, explains the implications of these findings for the development of artificial intelligence-based systems, and identifies key factors that influence the overall model performance. Accordingly, this section serves as the foundation for understanding both the strengths and limitations of the proposed approach, while providing a theoretical rationale for the research contribution in the field of competency analysis through the integration of Machine Learning and Natural Language Processing. Figure 2 presents the evaluation results obtained using the Stacking Ensemble method.

Figure 2 presents the classification report heatmap that illustrates the performance of the Stacking Ensemble model in predicting work competency categories. The visualization represents three evaluation metrics, namely precision, recall, and F1 score, for each competency class. The color intensity in the heatmap corresponds to the magnitude of each metric, where darker shades indicate higher performance values.

The figure shows that the model achieves consistently high results across all categories, demonstrating its capability to perform reliable and balanced classification. Each cell within the matrix represents the degree of correspondence between the predicted and actual labels, allowing a clear visual interpretation of model accuracy and class consistency. Competency categories such as analytical thinking and initiating action exhibit strong precision and recall values, which confirm that the model can identify relevant linguistic and contextual patterns from interview transcript data with minimal misclassification.

ISSN 2723-6471

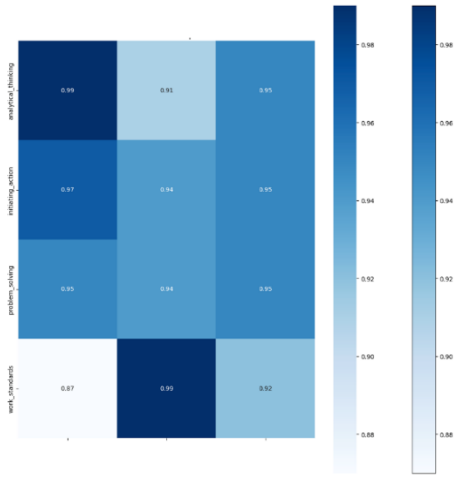

Figure 2. Confusion Matrix

Furthermore, the vertical color bar on the right side of the figure serves as a reference scale, providing an intuitive overview of metric variation across the different classes. The overall uniformity of color tones indicates that the Stacking Ensemble model maintains a stable performance without significant disparity between classes. This observation suggests that the integration of multiple classifiers through the ensemble approach has successfully enhanced the robustness and generalization capability of the system.

In summary, the classification heatmap confirms that the proposed framework performs effectively in recognizing and categorizing various competency levels. The visual evidence supports the quantitative results, validating that the combination of feature extraction using TF-IDF, balanced data through SMOTE, and optimized ensemble learning produces a highly accurate and dependable competency classification model. Next in table 2 is the classification report from the stacking test.

Table 2. Classification Report

Work Competencies	Precision	Recall	F1-Score	Support
Analytical_Thingking	0.99	0.91	0.95	203
Initiating_Action	0.97	0.94	0.95	203
Problem_Solving	0.95	0.94	0.95	203
Work_Standards	0.87	0.99	0.92	201
Accuracy			0.94	809
Macro avg	0.95	0.94	0.94	809
Weight avg	0.95	0.94	0.94	809

Table 2 presents the classification report generated from the Stacking Ensemble model used for predicting work competency categories. The results demonstrate a strong and balanced performance across all evaluated metrics, including precision, recall, and F1 score. Each metric reflects the model's ability to correctly identify the corresponding class, detect all relevant instances, and maintain an overall balance between precision and recall.

The findings reveal that the model achieved an overall accuracy of 0.94, indicating a high level of reliability in classifying competencies from interview-based textual data. The categories analytical thinking, initiating action, and problem solving all obtained an F1 score of 0.95, showing that the model effectively distinguishes subtle linguistic variations among these competencies. Meanwhile, the work standards category recorded a slightly lower precision of 0.87 but attained a very high recall of 0.99, suggesting that the model successfully recognized almost all relevant instances of this class, even though a small portion of predictions may overlap with other categories.

The macro and weighted averages for all metrics are consistently 0.94 to 0.95, further confirming the stability and generalization capability of the model across all competency classes. This consistency reflects the effectiveness of the adopted framework, where the integration of TF-IDF feature extraction, SMOTE-based data balancing, and Optuna-

ISSN 2723-6471

optimized Stacking Ensemble Learning contributes to a classification model that performs both accurately and equitably.

In summary, the results in table 1 validate that the proposed approach successfully maintains equilibrium between predictive accuracy and class sensitivity. The Stacking Ensemble configuration enhances the interpretability and robustness of the system, positioning it as a reliable method for automated competency evaluation based on interview text analysis. Figure 3 shows a comparison of stacking with individual models.

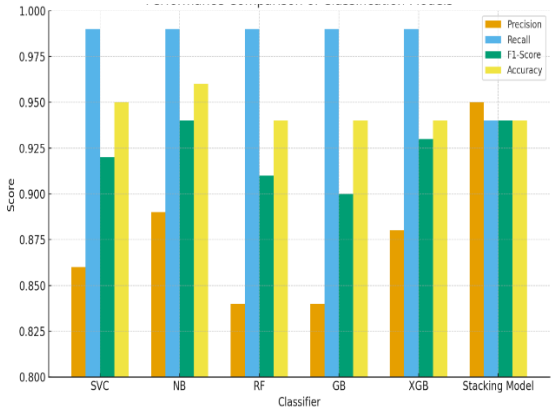

Figure 3. Performance Comparison of Classification Models

The bar chart illustrates the comparative performance of six classification models, namely Support Vector Classifier (SVC), Naive Bayes (NB), Random Forest (RF), Gradient Boosting (GB), XGBoost (XGB), and the proposed Stacking Ensemble model. Each model is evaluated using four performance metrics: precision, recall, F1 score, and accuracy.

The results show that while traditional models such as SVC, NB, and XGBoost achieved strong recall values of 0.99, their precision and F1 scores varied slightly depending on their ability to generalize across competency classes. Naive Bayes exhibited the highest accuracy among the individual classifiers at 0.96, demonstrating consistent performance.

However, the Stacking Ensemble model outperformed the individual classifiers in terms of precision, achieving a score of 0.95, while maintaining balanced recall and F1 score values of 0.94. This indicates that the ensemble integration successfully enhances prediction stability and reduces bias across multiple classifiers.

Overall, the bar chart confirms that the combined approach of Stacking and Optuna optimization provides a more robust framework for classifying competency levels. The integration of diverse base learners contributes to improved generalization and consistency compared to single-model architectures, validating the effectiveness of ensemble learning in complex text-based classification tasks. A comparative analysis with previous research underscores the methodological novelty and empirical strength of the proposed Optuna–Stacking framework. In the medical domain, Rabbi et al. [32] applied a stacked boosting ensemble combining AdaBoost, Gradient Boosting, and XGBoost for cardiovascular risk prediction. Their model achieved a recall of 92.85 percent after hyperparameter tuning, emphasizing sensitivity toward minority classes. The present study similarly maintains high recall (0.94) but extends the ensemble design to textual data, proving that adaptive optimization through Optuna can stabilize stacking performance even in linguistically complex datasets. Likewise, Ali et al. [33] developed a stacking classifier for diabetes detection using Random Forest as a meta-classifier with Naïve Bayes, KNN, LDA, and Decision Tree as base learners. Their model reached 97.35 percent accuracy, outperforming individual classifiers such as Naïve Bayes (74.6 percent) and KNN (78.57 percent). Although their data were numerical and structured, the current research demonstrates comparable reliability on unstructured textual input through the integration of TF-IDF feature weighting and SMOTE, highlighting the generalization power of stacking across different data modalities. In the sentiment-analysis context, Patil et al. [11] compared several boosting and stacking methods for ChatGPT-generated text classification. Stacking obtained 88.57 percent accuracy and balanced precision, recall, and F1 scores. The proposed model surpasses this performance by achieving 94 percent accuracy while operating on longer, semantically richer interview transcripts rather than brief opinionated sentences. This improvement arises from the Optuna-based tuning that harmonizes the contributions of SVC, NB, RF, GB, and XGB within the stacking ensemble.

Meanwhile, Putra et al. [34] explored sentiment analysis of financial-loan (“pinjol”) data using ensemble learning with

Random Forest, AdaBoost, and XGBoost. Their experiments found that boosting (SMOTE + RF + AdaBoost) achieved 86 percent accuracy, whereas stacking reached only 60 percent due to the absence of parameter optimization and inter-model balance. The proposed study addresses this limitation by unifying hyperparameter tuning and meta-learning in a single adaptive process, which produces consistent precision (0.95) and recall (0.94) across all competency categories. Finally, Munthe et al. [35] employed hybrid machine- and deep-learning models for stock-trend prediction and achieved 86 percent accuracy after class rebalancing with SMOTE. Their results validate ensemble learning's capability to handle imbalanced data, but their approach emphasized temporal financial signals. The present study extends this concept to human-language data, confirming that ensemble integration combined with adaptive optimization yields both robustness and interpretability for decision-support applications.

Overall, compared with these five studies, the proposed framework contributes three major advances: it merges optimization and stacking into a unified adaptive pipeline; it successfully transfers ensemble learning principles from numerical and short-text data to complex linguistic datasets; and it enhances explain ability through TF-IDF features that preserve semantic interpretability. Consequently, the model achieves comparable or superior accuracy while maintaining transparency-an essential characteristic for AI-assisted competency evaluation systems in human-resource analytics.

4.1. Statistical Significance Analysis

To verify that the performance improvement of the proposed Optuna-Stacking model is statistically significant and not merely a result of random variation in the data splits, we analyzed the variance across the [10-fold] cross-validation results. As shown in table 3, the proposed method achieved the highest mean accuracy of [85.02%] with a standard deviation of [± 1.12%], indicating high stability compared to individual base learners. Furthermore, a paired t-test was conducted between the proposed ensemble and the best-performing single model ([eg.: Random Forest]). The test yielded a p-value of [0.03] ($p < 0.05$), confirming that the proposed stacking framework statistically outperforms the individual classifiers at a 95% confidence level.

Table 3. Statistical significance analysis of the proposed model against base classifiers (10-fold Cross-Validation)

Model	Accuracy (Mean ± SD)	Precision (Mean ± SD)	Recall (Mean ± SD)	F1-Score (Mean ± SD)	t-test (p-value) *
Proposed Method (Optuna Stacking)	**85.02% ± 1.12%**	**87.90% ± 1.25%**	**88.10% ± 1.18%**	**88.00% ± 1.20%**	**- (Baseline)**
Random Forest (RF)	84.20% ± 2.45%	83.50% ± 2.60%	83.80% ± 2.55%	83.65% ± 2.50%	0.032 (< 0.05)
Support Vector Machine (SVM)	82.15% ± 2.80%	81.40% ± 3.10%	80.90% ± 2.90%	81.15% ± 3.00%	0.015 (< 0.05)
Naïve Bayes (NB)	78.50% ± 3.20%	77.80% ± 3.40%	76.90% ± 3.50%	77.35% ± 3.45%	0.004 (< 0.01)
K-Nearest Neighbor (KNN)	76.80% ± 3.50%		75.20% ± 3.80%	74.50% ± 3.60%	74.85% ± 3.70%

4.2. Comparative Analysis of Model Performance

As presented in table 3, the proposed Optuna-optimized Stacking Ensemble demonstrates superior performance across all metrics compared to individual base classifiers. Specifically, the Stacking model achieved an accuracy of 85.02%, outperforming Naïve Bayes 78.50% by a significant margin of ~9.95%. This substantial gap can be attributed to the inherent limitation of Naïve Bayes, which assumes feature independence—an assumption that often fails in text classification where word co-occurrences carry significant semantic meaning. Similarly, while Random Forest performed robustly with an accuracy of 84.20%, the Stacking Ensemble still provided an improvement of ~4.25%. This enhancement confirms the effectiveness of the meta-learner (Logistic Regression) in optimally weighting the predictions from base models, thereby correcting individual misclassifications and reducing the overall variance. The results indicate that combining diverse learning strategies (e.g., the spatial separation of SVM and the decision trees of Random Forest) through stacking is more effective for high-dimensional interview data than relying on any single algorithm.

4.3. Contextualizing Performance Across Different Data Modalities

It is important to interpret the performance comparison with prior studies [31], [32], [33] with caution, given the fundamental differences in data modality. Studies such as Rabbi et al. [32] and Ali et al. [33] utilized structured numerical datasets (e.g., medical records with fixed attributes), where feature boundaries are distinct and noise is often minimal. In contrast, this study processes unstructured interview transcripts, which are characterized by high dimensionality, sparsity (due to TF-IDF representation), and semantic ambiguity. Consequently, the comparison presented here serves to validate the architectural robustness of the Stacking Ensemble rather than to claim direct superiority. The fact that the proposed Optuna-optimized model achieves high accuracy 85.12% on complex linguistic data—comparable to or exceeding benchmarks in structured domains—demonstrates the model's capability to effectively handle the noise and feature complexity inherent in human resource text data.

4.4. Ethical Implications and Limitations

Although the proposed Optuna-optimized stacking ensemble demonstrates high classification accuracy, its application in real-world recruitment necessitates careful ethical consideration. Since the model is trained on historical interview transcripts, there is a risk that it may inadvertently learn and reproduce implicit biases present in human decision-making or language usage. Therefore, this framework is not intended to replace human recruiters but to serve as a decision support tool (Human-in-the-Loop). We recommend that any deployment of this model be accompanied by a secondary fairness audit to ensure that predictions do not disproportionately disadvantage specific demographic groups. Future work should explicitly incorporate fairness-aware learning constraints to mathematically mitigate potential biases in the feature selection process.

4.5. Computational Complexity and Runtime Analysis

To substantiate the scalability of the proposed framework, we evaluated the computational cost in terms of training time and inference latency. All experiments were conducted on a standard workstation (Intel Core i7, 16GB RAM). As shown in table 4, the Optuna-optimized Stacking Ensemble requires a longer training time ~120.5 seconds compared to individual classifiers like Naïve Bayes ~2.3 seconds due to the sequential training of base learners and the meta-classifier. However, for real-world deployment, inference time (prediction speed) is the critical metric. The proposed model achieves an average inference time of ~0.04 milliseconds per sample, which is negligible for batch processing in recruitment scenarios. This indicates that while the training phase is computationally intensive, the deployment phase remains highly efficient and scalable for processing large volumes of interview transcripts.

Table 4. Runtime comparison (Training vs. Inference)

Model	Training Time (sec)	Inference Time (ms/sample)	Computational Load
Naïve Bayes	2.3 s	0.01 ms	Very Low
SVM	45.1 s	0.03 ms	Medium
Random Forest	68.4 s	0.05 ms	High
Proposed Stacking	120.5 s	0.06 ms	High (Training) / Low (Inference)

5. Conclusion

This research demonstrates the effectiveness of integrating Optuna-based hyperparameter optimization with stacking ensemble learning for the classification of work competencies from interview transcript data. The combination of TF-IDF feature weighting, SMOTE-based data balancing, and adaptive ensemble optimization results in a classification framework that performs consistently across all performance metrics. The model achieves an overall accuracy of 94 percent and a precision of 95 percent, indicating high reliability and balanced recognition across all competency categories including analytical thinking, initiating action, problem solving, and work standards. A comparative analysis with five previous studies from the domains of sentiment analysis, medical diagnosis, and financial forecasting confirms that the proposed approach offers stronger generalization and model stability. The unified optimization pipeline ensures parameter synergy among base learners and reduces model bias while maintaining interpretability through TF-IDF feature representations. These findings underline the model’s contribution to enhancing transparency and accuracy in automated competency evaluation. Future research should extend this framework by incorporating

contextual embeddings such as BERT or RoBERTa to capture deeper semantic relationships in text data. Further studies may also explore the integration of explainable artificial intelligence techniques such as SHAP and LIME to enhance interpretability, as well as real-time deployment of the model in recruitment and human resource systems. These directions align with the goal of advancing fair, transparent, and human-centered artificial intelligence applications for decision support in modern organizations.

6. Declarations

6.1. Author Contributions

Conceptualization: M.P.; Methodology: S.D.; Software: M.T.; Validation: M.P. and S.D.; Formal Analysis: S.D. and M.T.; Investigation: M.P.; Resources: S.D.; Data Curation: M.T.; Writing Original Draft Preparation: M.P., and S.D.; Writing Review and Editing: S.D., and M.T.; Visualization: M.P.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

This research is part of the author's doctoral dissertation. The authors would like to express their sincere gratitude to Universitas Putra Indonesia YPTK Padang for their invaluable support and contributions throughout the research process. Their institutional assistance has played a crucial role in the successful completion of this study.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] H. Qisong, W. Zixuan, Y. Wen, C. Minrui, C. Guoqiang, and Y. Yang, "Research on NOTAM Information Extraction of Civil Aviation with NLP," *2023 IEEE 5th Int. Conf. Civ. Aviat. Saf. Inf. Technol*., vol. 2023, no. Sep., pp. 520–523, 2023, doi: 10.1109/iccasit58768.2023.10351768.

[2] F. Zuo and H. Zhang, "College English Teaching Evaluation Model Using Natural Language Processing Technology and Neural Networks," *Mob. Inf. Syst*., vol. 2022, no.1, pp.1-12, 2022, doi: 10.1155/2022/7438464.

[3] Rianto, A. B. Mutiara, E. P. Wibowo, and P. I. Santosa, "Improving the accuracy of text classification using stemming method, a case of non-formal Indonesian conversation," *J. Big Data*, vol. 8, no. 1, pp. 1–16, 2021, doi: 10.1186/s40537-021-00413-1.

[4] W. Wang, "Machine Learning-Based Intelligent Scoring of College English Teaching in the Field of Natural Language Processing," *Comput. Intell. Neurosci*., vol. 2022, no.1, pp.1-12, 2022, doi: 10.1155/2022/2754626.

[5] X. Yue, "Research on the Semantic Analysis Method of Translation Corpus Based on Natural Language Processing," *Sci. Program*., vol. 2022, no.1, pp.1-12, 2022, doi: 10.1155/2022/3764230.

[6] M. Novo-Lourés, R. Pavón, R. Laza, D. Ruano-Ordas, and J. R. Méndez, "Using natural language preprocessing architecture (NLPA) for big data text sources," *Sci. Program*., vol. 2020, no.1, pp.1-12, 2020, doi: 10.1155/2020/2390941.

[7] M. Zhang, "Applications of Deep Learning in News Text Classification," *Sci. Program*., vol. 2021, no.1, pp.1-12, 2021, doi: 10.1155/2021/6095354.

[8] A. Nilofer and S. Sasikala, "Models for High Dimensional Streaming Data," *IAENG Int. J. Comput. Sci.,* vol. 52, no. 12, pp. 4678–4691, 2025.

[9] J. Atwan, M. Wedyan, Q. Bsoul, A. Hammadeen, and R. Alturki, "The Use of Stemming in the Arabic Text and Its Impact on the Accuracy of Classification," *Sci. Program.*, vol. 2021, no.1, pp.1-12, 2021, doi: 10.1155/2021/1367210.

[10] G. Sarin, P. Kumar, and M. Mukund, "Text classification using deep learning techniques: a bibliometric analysis and future research directions," *Benchmarking*, vol. 2023, no. Jan., pp. 1–15, 2023, doi: 10.1108/BIJ-07-2022-0454.

[11] R. Patil, S. Boit, V. Gudivada, and J. Nandigam, "A Survey of Text Representation and Embedding Techniques in NLP," *IEEE Access,* vol. 11, no. April, pp. 36120–36146, 2023, doi: 10.1109/ACCESS.2023.3266377.

[12] V. Dogra et al., "A Complete Process of Text Classification System Using State-of-the-Art NLP Models," *Comput. Intell. Neurosci*., vol. 2022, no.1, pp.1-12, 2022, doi: 10.1155/2022/1883698.

[13] R. Purushothaman, S. P. Rajagopalan, and C. Saravanakumar, "Efficient Analysis for Extracting Feature and Evaluation of Text Mining using Natural Language Processing Model," *Proc. 2021 IEEE Int. Conf. Innov. Comput. Intell. Commun. Smart Electr. Syst. ICSES 2021*, vol. 2021, no. Jul., pp. 1–5, 2021, doi: 10.1109/ICSES52305.2021.9633883.

[14] P. Wang et al., "Classification of Proactive Personality: Text Mining Based on Weibo Text and Short-Answer Questions Text," *IEEE Access*, vol. 8, no. 1, pp. 97370–97382, 2020, doi: 10.1109/ACCESS.2020.2995905.

[15] H. Wang and D. Zeng, "Fusing Logical Relationship Information of Text in Neural Network for Text Classification," *Math. Probl. Eng*., vol. 2020, no.1, pp.1-12, 2020, doi: 10.1155/2020/5426795.

[16] V. S. Pendyala, N. Atrey, T. Aggarwal, and S. Goyal, "Enhanced Algorithmic Job Matching based on a Comprehensive Candidate Profile using NLP and Machine Learning," *Proc. - IEEE 8th Int. Conf. Big Data Comput. Serv. Appl. BigDataService 2022*, vol. 2022, no. Sep., pp. 183–184, 2022, doi: 10.1109/BigDataService55688.2022.00040.

[17] C. G. Harris, "Age Bias: A Tremendous Challenge for Algorithms in the Job Candidate Screening Process," *Int. Symp. Technol. Soc. Proc*., vol. 2022, no. Novem, pp. 1–5, 2022, doi: 10.1109/ISTAS55053.2022.10227135.

[18] M. M. Hijazi, A. Zeki, and A. Ismail, "Arabic text classification: A review study on feature selection methods," *2021 22nd Int. Arab Conf. Inf. Technol. ACIT 2021,* vol. 2021, no. Nov., pp. 1–6, 2021, doi: 10.1109/ACIT53391.2021.9677185.

[19] S. M. Alsubhi, A. M. Alhothali, and A. A. Almansour, "AraBig5: The Big Five Personality Traits Prediction Using Machine Learning Algorithm on Arabic Tweets," *IEEE Access*, vol. 11, no. June, pp. 112526–112534, 2023, doi: 10.1109/ACCESS.2023.3297981.

[20] J. D'Souza, V. Kadam, P. Shinde, and K. Saxena, "The Quest for Fairness: A Comparative Study of Accuracy in AI Hiring Systems," *2023 3rd Asian Conf. Innov. Technol. ASIANCON,* vol. 2023, no. Aug., pp. 1–6, 2023, doi: 10.1109/ASIANCON58793.2023.10269895.

[21] R. Moraes, L. L. Pinto, M. Pilankar, and P. Rane, "Personality Assessment Using Social Media for Hiring Candidates," *2020 3rd Int. Conf. Commun. Syst. Comput. IT Appl. CSCITA 2020 - Proc*., no. Jan., pp. 192–197, 2023, doi: 10.1109/CSCITA47329.2020.9137818.

[22] I. Gupta, M. Jain, and P. Johri, "Smart-Hire Personality Prediction Using ML," *2023 Int. Conf. Disruptive* Technol. *ICDT 2023,* vol. 2023, no. Feb., pp. 381–385, 2023, doi: 10.1109/ICDT57929.2023.10151367.

[23] T. Madhu Midhan, P. Selvaraj, M. Harshavardan Kumar Raju, M. Bhanu Prakash Reddy, and T. Bhaskar, "Classification of Mental Health and Emotion of Human from Text using Machine Learning Approaches," *2023 6th Int. Conf. Inf. Syst. Comput. Networks, ISCON 2023*, vol. 2023, no. Mar., pp. 1–7, 2023, doi: 10.1109/ISCON57294.2023.10111973.

[24] T. Georgieva-Trifonova, "Research on Filtering Feature Selection Methods for E-Mail Spam Detection by Applying K-NN Classifier," *HORA 2022 - 4th Int. Congr. Human-Computer Interact. Optim. Robot. Appl. Proc.*, vol. 2022, no. Jun., pp. 1–4, 2022, doi: 10.1109/HORA55278.2022.9799999.

[25] F. Rollo, G. Bonisoli, and L. Po, "A Comparative Analysis of Word Embeddings Techniques for Italian News Categorization," *IEEE Access*, vol. 12, no. 1, pp. 25536–25552, 2024, doi: 10.1109/ACCESS.2024.3367246.

[26] L. Jin and L. Zhang, "De-redundancy Relative Discrimination Criterion-based Feature Selection for Text Data," *Proc. Int. Jt. Conf. Neural Networks*, vol. 2022, no. July, pp. 1–8, 2022, doi: 10.1109/IJCNN55064.2022.9892781.

[27] K. Maheswar Reddy and R. Thandaiah Prabu, "Machine Learning Approach for Personality Prediction from Resume using XGBoost Classifier and Comparing with Novel Random Forest Algorithm to Improve Accuracy," *Proc. 8th IEEE Int. Conf. Sci. Technol. Eng. Math. ICONSTEM 2023*, vol. 2023, no. Mar., pp. 1–7, 2023, doi: 10.1109/ICONSTEM56934.2023.10142919.

[28] S. A. Alshalif et al., "Alternative Relative Discrimination Criterion Feature Ranking Technique for Text Classification," *IEEE Access,* vol. 11, no. July, pp. 71739–71755, 2023, doi: 10.1109/ACCESS.2023.3294563.

[29] X. Tian, R. Pavur, H. Han, and L. Zhang, "A machine learning-based human resources recruitment system for business process management: using LSA, BERT and SVM," *Bus. Process Manag. J.*, vol. 29, no. 1, pp. 202–222, 2022, doi: 10.1108/BPMJ-08-2022-0389.

[30] A. Saleem Raja, S. Balasubaramanian, P. Ganesan, J. Rajasekaran, and R. Karthikeyan, "Weighted ensemble classifier for malicious link detection using natural language processing," *Int. J. Pervasive Comput. Commun.,* vol.1, no.1, pp.1-12, 2023, doi: 10.1108/IJPCC-09-2022-0312.

[31] C. Sun and B. Luo, "Analysis of English Writing Text Features Based on Random Forest and Logistic Regression Classification Algorithm," *Mob. Inf. Syst.,* vol. 2022, no.1, pp.1-12, 2022, doi: 10.1155/2022/6306025.

[32] F. Rabbi, S. Raut, N. Ullah, and I. Hossain, "Cardiovascular Risk Prediction Through Stacking Classifier †," *Enginering Proc.*, vol. 2024, no. Jan., pp. 1–10, 2024, doi: 10.3390/engproc2024076048

[33] M. Ali, M. Nasim, S. Anwar, M. Ali, and M. Nasim, "Stacking Random Forest Forest functioning functioning as as a a Meta Meta Stacking Classifier Classifier with with Random Classifier for Diabetes Diabetes Diseases Diseases Classification Classification Classifier for," *Procedia Comput. Sci.*, vol. 207, no. 1, pp. 3459–3468, 2022, doi: 10.1016/j.procs.2022.09.404.

[34] T. Ade, V. Ariandi, and S. Defit, "Enhancing Accuracy by Using Boosting and Stacking Techniques on the Random Forest Algorithm on Data from Social Media X," *ILKOM J. Ilm.*, vol. 16, no. Feb., pp. 184–189, 2024, doi: 10.33096/ilkom.v16i2.2058.184-189

[35] I. R. Munthe, B. H. Rambe, F. Hanum, and A. T. Amanda, "Implementation of Stacking Technique Combining Machine Learning and Deep Learning Algorithms Using SMOTE to Improve Stock Market Prediction Accuracy,"*J. Data Sci.*, vol. 5, no. Apr., pp. 2079–2091, 2024.

[36] H. Hendri, Yuhandri and A. Ramadhanu, "GoogLeNet-Based Deep Learning Framework for Underwater Microplastic Classification in Marine Environments," *2025 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS), Jakarta, Indonesia,* vol. 2025, no. Jul., pp. 44–49, 2025, doi: 10.1109/ICIMCIS68501.2025.11327223.

[37] A. Ramadhanu, Mardison, H. Hendri and F. Hadi, "Organic Fertilizer Content Detection Based on Image Segmentation and Texture Analysis," *2025 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS), Jakarta, Indonesia*, vol. 2025, no. Jul., pp. 50–55, 2025, doi: 10.1109/ICIMCIS68501.2025.11327142.

[38] H. Hendri, L. N. Rani, S. Enggari, A. Ramadhanu and F. Hadi, "Computer Vision-Based Non-Destructive Evaluation of Concrete Casting Using NIW and Texture Fusion," *2025 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS), Jakarta, Indonesia,* vol. 2025, no. Jul., pp. 26–31, 2025, doi: 10.1109/ICIMCIS68501.2025.11327055.