

# Optimization of SVM and Gradient Boosting Models Using GridSearchCV in Detecting Fake Job Postings

Rofik<sup>1</sup>, Roshan Aland Hakim<sup>1</sup>, Jumanto<sup>1</sup>, Budi Prasetyo<sup>1</sup>, Much Aziz Muslim<sup>2</sup>

<sup>1</sup>Universitas Negeri Semarang, Semarang, Indonesia

<sup>2</sup>Universiti Tun Hussein Onn Malaysia, Batu Pahat, Malaysia

---

## Article Info

### Article history:

Received November 15, 2023

Revised January 19, 2024

Accepted March 15, 2024

### Keywords:

*Fake Job Recruitment*

*Fraud Detection*

*Gradient Boosting*

*GridSearchCV*

*Support Vector Machine*

## ABSTRACT

Online job searching is one of the most efficient ways to do this, and it is widely used by people worldwide because of the automated process of transferring job recruitment information. The easy and fast process of transferring information in job recruitment has led to the rise of fake job vacancy fraud. Several studies have been conducted to predict fake job vacancies, focusing on improving accuracy. However, the main problem in prediction is choosing the wrong parameters so that the classification algorithm does not work optimally. This research aimed to increase the accuracy of fake job vacancy predictions by tuning parameters using GridSearchCV. The research method used was SVM and Gradient Boosting with parameter adjustments to improve the parameter combination and align it with the predicted model characteristics. The research process was divided into preprocessing, feature extraction, data separation, and modeling stages. The model was tested using the EMSCAD dataset. This research showed that the SVM algorithm can achieve the highest accuracy of 98.88%, while gradient enhancement produces an accuracy of 98.08%. This research showed that optimizing the SVM model with GridSearchCV can increase accuracy in predicting fake job recruitment.

Copyright ©2024 The Authors.

This is an open access article under the [CC BY-SA](#) license.



---

## Corresponding Author:

Jumanto, +6281339762820

Department of Computer Science, Faculty of Mathematics and Natural Sciences,

Universitas Negeri Semarang, Semarang, Indonesia,

Email: [jumanto@mail.unnes.ac.id](mailto:jumanto@mail.unnes.ac.id).

---

## How to Cite:

R. Rofik, R. A. Hakim, J. Unjung, B. Prasetyo, and M. A. Muslim, "Optimization of SVM and Gradient Boosting Models Using GridSearchCV in Detecting Fake Job Postings", *MATRIK: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer*, Vol. 23, No. 2, pp. 419-430, Mar, 2024.

This is an open access article under the CC BY-SA license (<https://creativecommons.org/licenses/by-sa/4.0/>)

## 1. INTRODUCTION

Technological developments like today, coupled with the emergence of artificial intelligence, machine learning, and so on, make almost all jobs such as education, health, government, business, and others be completed more easily [1–3], especially with internet media [4]. This is because information traffic moves very fast in this digital era. Like online job search, this is one of the most efficient ways that many people worldwide use because of the automatic job recruitment information transfer process [5, 6]. Companies do not need to spend too much money to announce job recruitment information, and so do prospective job applicants who do not need to spend more money, time, and effort to obtain this information [7, 8]. However, due to the rapid growth of information on the internet today, the world can easily deal with the problem of misinformation, such as fake content that comes in several forms and how it is packaged [9]. So, the ability to assess the credibility of information on the internet is considered an important topic by various fields of study, such as information science, psychology, sociology, and digital technology [10].

Many job recruitment offers on the Internet, and easy access to information that can be done anytime and anywhere are major advantages in the job search process. However, with the ease of accessing various information on the internet, the risk of fraud also increases, both in the field of job recruitment [11]. Instead of being an opportunity to share job recruitment information quickly, it has now increased the number of fake job recruitment, which annoys many people [12]. Job recruitment fraud on the internet is a dangerous act that damages the reputation of stakeholders, steals personal information, and causes economic loss [13].

Job recruitment or employment scams have increased during the Covid-19 pandemic. According to Consumer News and Business Channel (CNBC), job scams have doubled since 2017 [14]. As a result, many victims are deceived by fake job recruitment because fraudsters provide very tempting job offers to job seekers [15]. Scams can occur in various ways: (1) Job advertisements with attractive amounts of income to collect applicant personal data, including address, bank account, social security number, and more (2) Asking applicants to take an online test for a few minutes and then direct them to fake application sites to collect banking information, payments for specific purposes, and (3) downloading any virus or malware on the applicant's computer or mobile device to obtain survey system history [14]. This makes it necessary to detect fake job recruitment on the internet for the safety of many people. Machine learning is a form of technological development that is very helpful in automatically overcoming fraud patterns, as in this case in classifying or predicting [16, 17].

However, previous research has conducted fraud detection in job recruitment using XGBoost. The research managed to get an accuracy of 97.94% [13]. The research found that the organizational type of feature is the best feature in detecting recruitment fraud as an independent model. Research conducted by Naude et al. also focuses on detecting fake job recruitment types using the Gradient Boosting algorithm and using the steamy empirical rule set feature part-of-speech tags and bag-of-words vectors [18]. The research succeeded in achieving an F1 score of 88%. Research related to the detection of fake job recruitment has also been carried out by Bandyopadhyay et al., which used single and ensemble classifiers; this study found that the Random Forest algorithm achieved an accuracy of 98.27% [12]. SVM and XGBoost in detecting fake job recruitment have also been carried out before and produced a quite good performance with an accuracy rate of 87.4% and 98.53 [19]. The FJD-OT technique, namely the oversampling technique to increase predictability in detecting fake job recruitment, removal of stopwords, tokenizer, TF-IDF, and the application of SVM SMOTE to balance fake and real data has been carried out by previous studies and showed an increase in predictability [20]. By applying the Bi-LSTM model, the previous research also obtained quite good performance with an accuracy of 98.71% in detecting genuine and fake job recruitment [21].

So far, previous studies have successfully utilized various methods to detect fraud in job recruitment. Although the results obtained in previous studies have achieved good accuracy, some gaps can still be developed to optimize the accuracy obtained. Therefore, this research seeks to make further contributions in focusing on developing fake job recruitment detection models, specifically by integrating SVM and Gradient boosting algorithms using hyperparameter tuning with GridSearchCV. With the new methods used, this research aims to fill the gaps that still exist in the literature and improve the accuracy of fake job recruitment detection to a higher level.

The remainder of this paper is organized as follows: Section 2 provides a research framework for feature extraction, SVM-Gradient boosting model, parameter tuning using GridSearchCV, and model evaluation. Then, Section 3 compares the results of these experiments. Finally, we conclude the paper by summarizing our contributions and discussing future research directions in Section 4.

## 2. RESEARCH METHOD

This research was carried out in stages. The research phase is divided into preprocessing, feature extraction, split data, and Modeling. The research steps carried out can be seen in Figure 1.

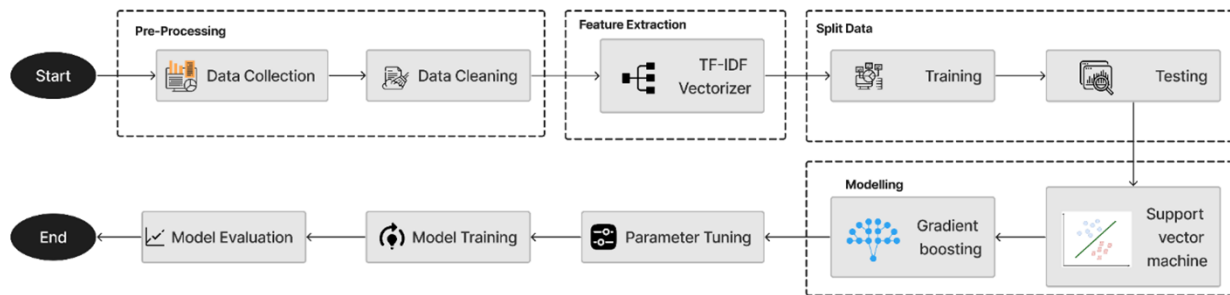


Figure 1. Research Method

## 2.1. Pre-Processing

The first stage in pre-processing is data collection and data cleaning. This research uses a public dataset from Kaggle to predict fake job recruitment. The dataset used is EMSCAD (Employment Scam Aegean Dataset), which can be accessed at the URL link: <https://www.kaggle.com/datasets/amruthjithrajvr/recruitment-scam>. The dataset contains 17,880 job recruitment ads. The features in the dataset include 18 features. The features include title, location, department, salary range, company profile, description, requirements, benefits, telecommuting, the company logo, questions, employment type, required experience, required education, industry, function, and fraud in the balanced dataset. The dataset has been labeled with 17,014 legitimate job ads and 866 fake job recruitment ads. The job recruitment ads were published from 2012 to 2014. After the data is collected, the next step is to clean the data. The data is loaded and processed using the Python programming language on the Google Collaboratory platform. Cleaning data is crucial in preparing and ridding the data of various attributes that may interfere with the modeling performance [22]. Data cleaning includes the process of identifying, correcting, and fixing errors, inconsistencies, or discrepancies in the data, such as deleting unnecessary columns, filling empty values with empty strings, and merging multiple columns into a single 'text' column. Data cleaning ensures that the data used in analysis or modeling is high quality, reliable, and clean. High-quality data creates a solid basis for rational decisions and reliable results in data analysis.

## 2.2. Feature Extraction

Feature Extraction was performed in this study to convert the job description text into a numerical representation for model analysis and training. Recruitment detection TF-IDF (Term Frequency-Inverse Document Frequency) is used in fake jobs. TF-IDF is a statistical method that performs calculation in the form of multiplying two metrics in a set of text, then dividing it by the number of occurrences of a word in a document (TF) and the inverse document frequency (IDF) of the word [23]. In this case, TF-IDF helps illustrate the importance of words that appear more frequently in job descriptions. It also helps highlight the keywords in the job posting content that can differentiate between fake and genuine recruitment. TF-IDF also helps in selecting features that are significant in distinguishing between fake and real job recruitment classes through word weighting [24, 25]. Words that appear frequently in the data content are given high weights, while low weights for words common in documents or job categories. Not only that, but TF-IDF in this classification also helps in dimensionality reduction or feature reduction. Because the analysis of datasets in the form of text often experiences high dimensionality due to the large number of words. So, words that have low weight or are less important can be removed or given lower weight, thus reducing the dimensionality of the features used in modeling. Thus, the application of feature extraction is very important in research.

## 2.3. Data Splitting

In this step, the dataset is divided into a subset of training data (training data) and a subset of test data (testing data). This separation is important for testing model performance on data that has never been seen. A comparison between training and test data is between 80% and 20%. Split data is done using the library in sklearn using the train test split function.

## 2.4. Model Building

The modeling process is performed using the SVM (Support Vector Machine) algorithm and the Gradient Boosting algorithm. This process aims to develop a fake job recruitment classification model to achieve high accuracy. The dataset used in this research contains a lot of text in the form of descriptions. Furthermore, this SVM algorithm is used because SVM is often used in conjunction with text datasets [26]. The concept of this algorithm is to focus on separating two classes with a line called a hyperplane with a maximum margin, which is the largest distance between the hyperplane and the closest points of each class of data. The decision obtained between the two classes is also maximized with the maximum margin. The advantage of this algorithm is that if the data used cannot be separated linearly, then non-linear transformation techniques can be performed to map the data to a higher dimension. SVM also uses a kernel function to calculate the distance between data points, which can streamline SVM performance. Thus, SVM is suitable for detecting job recruitment between fake and real [27]. Another advantage of this SVM algorithm is the ability to reduce overfitting and the ability to use many features [28]. Here is the modeled formula for the operation of SVM with a linear kernel in Equation (1) and RBF (Gaussian) in Equation (2).

$$f(x) = \text{sign}\left(\sum_{i=1}^N a_i y_i (x^T y_i) + b\right) \quad (1)$$

where :

$a_i$  is the Lagrange coefficient, determining the importance of each training sample in the model.  $y_i$  is the class label of the training sample, helping determine the direction and polarity of the decision.  $x$  is the input feature vector to be predicted and tested for its proximity to the hyperplane.  $b$  is the bias parameter, determining the hyperplane's location relative to the data's center.

In the case of a linear kernel, SVM attempts to create a straight line (hyperplane) to separate data between classes. This function examines how close each test data point ( $x$ ) is to this line. If the result is positive, the data is considered class one; otherwise, if negative, it is considered another class. The SVM formula using the RBF kernel is as given in Equation (2).

$$f(x) = \text{sign}\left(\sum_{i=1}^N a_i y_i \exp\left(-\frac{\|x - y_i\|^2}{2\sigma^2}\right) + b\right) \quad (2)$$

where :

$\sigma$  is a parameter controlling the sharpness of the RBF kernel function curve. A larger value makes the function flatter, while a smaller value makes it sharper. This influences how sensitive the model is to the distance between data points.

In the case of an RBF kernel, SVM operates by evaluating how close the test data points ( $x$ ) are to the training points ( $y_i$ ) using an exponential function. A positive result indicates class one, while a negative result indicates another class.

In addition, the Gradient Boosting algorithm is also modeled in this research. Gradient Boosting is an ensemble learning algorithm that creates a strong model with high accuracy by combining many weak learnings with relatively low accuracy [29, 30]. This algorithm works by iteratively learning and improving the weaknesses of previous models by emphasizing samples that are difficult to classify. The process starts with predicting the target variable, and then the model is built to shape the residuals generated by the previous model. The process is repeated while optimizing the model parameters and weighting each model's performance to correct previous prediction errors. The advantage of this algorithm is that it handles class imbalance, and it is suitable due to the different numbers of legitimate jobs and fake jobs in the dataset used. Here is how gradient boosting works. Start by initializing predictions using a simple decision tree, as shown in Equation (3), and calculate the residual resulting from predictions as in Equation (4).

$$f_0(x) = \text{argim} \sum_{i=1}^n L(y_i, y) \quad (3)$$

$$\tilde{y}_{im} = -\left[\frac{\partial \Psi(y_i, F(x_i))}{\partial F(x_i)}\right] = F_{m-1} \quad (4)$$

Build an additional decision tree that predicts the residual values from all independent variables. Update predictions with newly calculated values using the learning rate, as in Equation 5.

$$F_m(x) = F_{m-1}(x) + v \sum_{j=1}^J m y_{jm} 1(x \in R_{jm}) \quad (5)$$

Iterate again from steps 2 to 4 according to the specified number of iterations or the predetermined number of trees; the two algorithms, SVM and Gradient Boosting, produce a model that can be used to predict and detect fake job recruitment.

## 2.5. Parameter Tuning

In the process of developing SVM and Gradient Boosting algorithm models, parameter tuning is conducted to find optimal parameter combinations, ensuring that the built classification models perform optimally in detecting fake job recruitments. This research utilizes the GridSearchCV method, which evaluates models using possible combinations from a parameter grid. It explores all specified parameter combinations through cross-validation. The SVM algorithm's parameters include the 'C' parameter controlling the trade-off between maximum margin and misclassification quantity. Tuning is also performed on the kernel to map data to higher dimensions, and the 'gamma' parameter controls the extent of the influence of sample data. Parameters 'C' with values [0.1, 1, 10], 'kernel' with values ['linear', 'rbf'], and 'gamma' with values ['scale', 'auto'] are evaluated and compared to discover the optimal parameter combination. On the other hand, the Gradient Boosting algorithm employs several tuning parameters such as 'n\_estimator,' determining the number of decision trees to be built and used in the ensemble. These trees are constructed sequentially, with each tree attempting to correct prediction errors made by the previous tree. The 'learning\_rate' parameter controls the contribution of each tree in the ensemble, with smaller values requiring more trees to build a robust model. The 'max\_depth' parameter measures the maximum depth of each tree in the ensemble, as excessively high values may cause overfitting. At the same time, values that are too low may result in a model that is too simplistic and unfit. Parameters 'n\_estimator' with values [50, 100], 'learning\_rate' with values [0.1, 0.5], and 'max\_depth' with values [3, 5] are also evaluated and compared to find the optimal parameter combination for the Gradient Boosting model's best performance.

On the other hand, the Gradient Boosting algorithm employs several tuning parameters such as 'n\_estimator,' determining the number of decision trees to be built and used in the ensemble. These trees are constructed sequentially, with each tree attempting to correct prediction errors made by the previous tree. The 'learning\_rate' parameter controls the contribution of each tree in the ensemble, with smaller values requiring more trees to build a robust model. The 'max\_depth' parameter measures the maximum depth of each tree in the ensemble, as excessively high values may cause overfitting. At the same time, values that are too low may result in a model that is too simplistic and unfit. Parameters 'n\_estimator' with values [50, 100], 'learning\_rate' with values [0.1, 0.5], and 'max\_depth' with values [3, 5] are also evaluated and compared to find the optimal parameter combination for the Gradient Boosting model's best performance.

## 2.6. Model Evaluation

Model evaluation tests the model constructed using previously unseen test data. This evaluation is performed to measure the model's performance through accuracy (6), precision (7), recall (8), and F1-score (9) values. Model evaluation is carried out using a confusion matrix table to assess the performance of each algorithm [31]. The confusion matrix is a table that provides information about the number of correct and incorrect predictions for each class. From this confusion matrix, we can assess the classification model's performance by comparing the predictions of each algorithm with the actual values. The following is the formula to calculate the model performance.

### 1. Accuracy

Accuracy is a value that indicates how accurate the model is in predicting the entire data. The formula for calculating accuracy can be seen in Equation (6):

$$Accuracy = \frac{(TP + TN)}{TP + TN + FP + FN} \quad (6)$$

TP (True Positive): In this case, the model correctly detects the number of fake jobs.

TN (True Negative): Indicates that the model correctly detected the original work.

FP (False Positive): This indicates the number of original jobs, but the model classified it as fake.

FN: Indicates the number of bogus jobs but is found to be classified as genuine by the model.

## 2. Precision

Precision measures the degree to which work predicted to be fake is fake. The formula for calculating precision is in Equation (7).

$$Precision = \frac{TP}{(TP + FP)} \quad (7)$$

## 3. Recall (Sensitivity)

Recall measures the extent to which the model successfully detects fake jobs overall. The formula for calculating recall is in Equation (8).

$$Recall = \frac{TP}{(TP + FN)} \quad (8)$$

## 4. F1-Score

The F1-Score combines precision and recall into a single metric that yields the overall model performance. The formula for calculating the F1-Score is in Equation (9).

$$F1 - Score = \frac{2 * (Presisi * Recall)}{(Presisi + Recall)} \quad (9)$$

### 3. RESULT AND ANALYSIS

The process of classifying fake job vacancies begins by collecting data sourced from Kaggle, namely The Employment Scam Aegean Dataset (EMSCAD), with a record number of 17880, of which 866 records are fake job vacancies and 17014 records are genuine job vacancies. Data is processed using Python with a tool called Colab. The dataset consists of 18 features, which include the following attributes: title, location, department, salary\_range, company\_profile, description, requirements, benefits, telecommuting, has\_company\_logo, has\_questions, employment\_type, required\_experience, required\_education, industry, function, fraudulent, in\_balanced\_dataset. When the dataset has been obtained, it enters the pre-processing and modeling stages, which are carried out in the Colab itself. Data is cleaned by deleting empty data and selecting features. After the data cleaning process is completed, the data is divided into training and testing data. In comparison, the split is 80% and 20%. A total of 14,304 training data and 3,576 testing data. The data is trained using the SVM algorithm and the Gradient Boosting algorithm. There are very important features in the dataset that support the classification of fake job vacancies in this study. These features are company\_profile, description, requirements, and benefits.

Extraction is carried out with the outfitter using TF-IDF (Term Frequency - Inverse Document Frequency) to maximize the quality of the results from SVM and Gradient Boosting modeling results. This is important for tackling common words that do not provide relevant information and identifying words that are important in distinguishing genuine from fake work. From the modeling implementation of the two algorithms, namely SVM and Gradient Boosting, the model's accuracy can be seen in Figure 2.

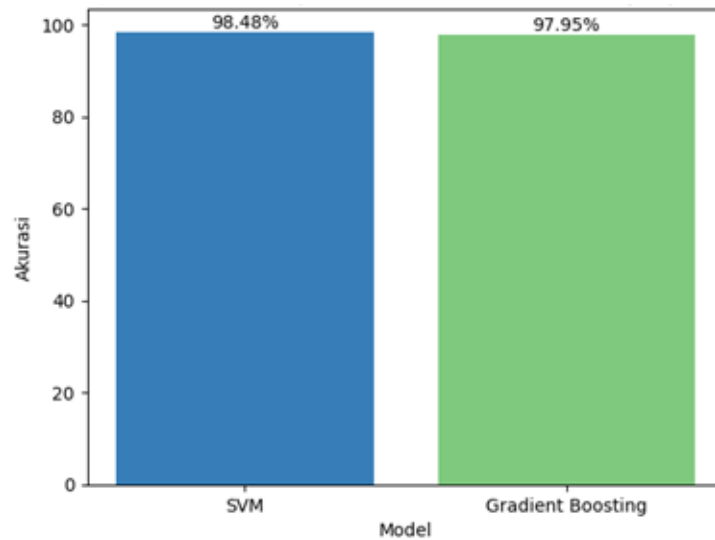


Figure 2. Comparison of model accuracy.

Parameter tuning is carried out using GridSearchCV to improve the performance of the algorithm being modeled. By trying parameter combinations carried out by the GridSearchCV method as a whole, it is possible to find parameters that match the predicted model characteristics [32]. Parameter settings calculated by GridSearchCV for the SVM algorithm are listed in Table 1.

Table 1. Hyperparameters for the SVM Model

Tuning Hyperparameter	Values and Ranges	Optimal Hyperparameter
C	[0.1, 1, 10]	10
kernel	[linier, rbf]	linier
gamma	[scale, auto]	scale

GridSearchCV tries all possible combinations of parameter values of param\_grid and evaluates its performance using an evaluation metric. In this study, the accuracy score is used as the evaluation metric. A cross-validation of 5 folds was also performed.

The parameter settings calculated by GridSearchCV for the Gradient Boosting algorithm are listed in Table 2.

Table 2. Hyperparameters for Gradient Boosting Model

Tuning Hyperparameter	Values and Ranges	Optimal Hyperparameter
n_estimators	[50, 100]	100
learning_rate	[0.1, 0.5]	0.1
max_depth	[3, 5]	5

From the results after parameter tuning, the algorithm model showed better performance, as seen in Figure 3.

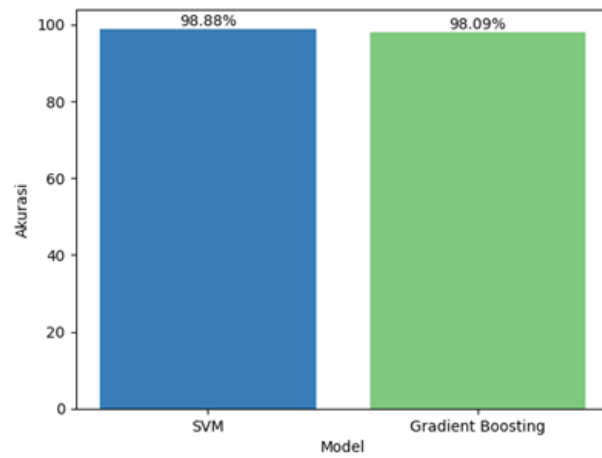


Figure 3. Comparison of model accuracy after parameter tuning

Figures 4, 5, 6 and 7 show the results of evaluating the SVM and Gradient Boosting algorithm models on the confusion matrix before and after parameter tuning.

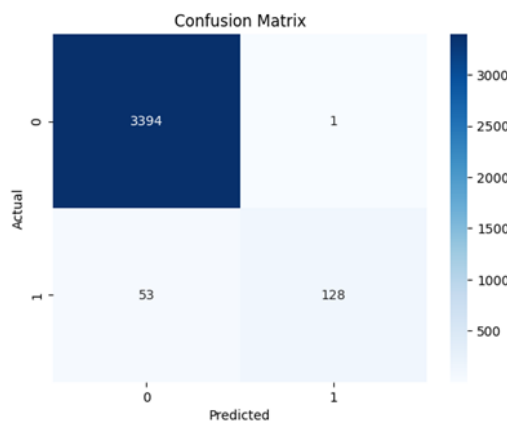


Figure 4. Confusion Matrix SVM Algorithm

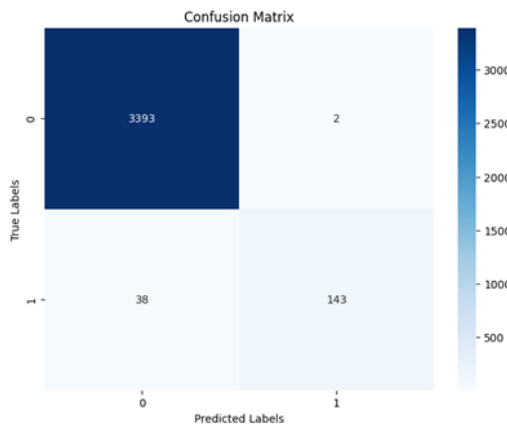


Figure 5. Confusion Matrix of SVM Algorithm after Parameter Tuning



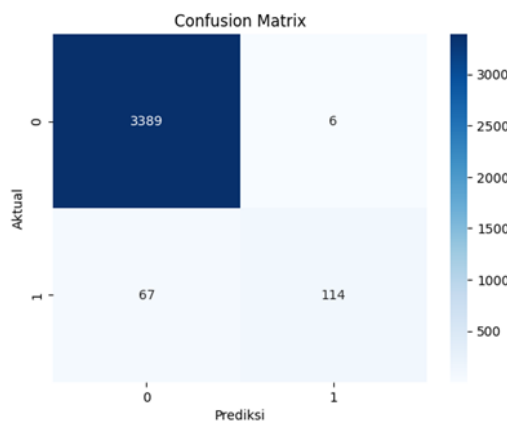


Figure 6. Confusion Matrix of Gradient Boosting Algorithm

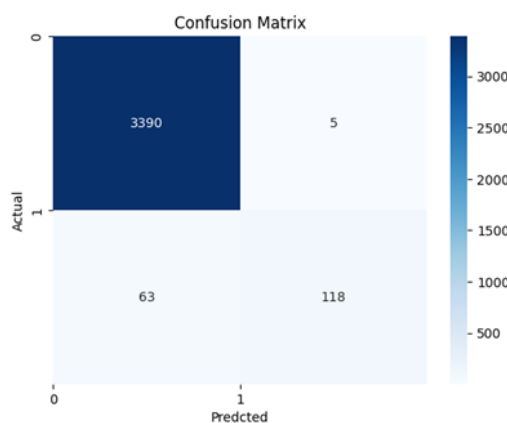


Figure 7. Confusion Matrix of Gradient Boosting Algorithm after Parameter Tuning

Table 3 shows the confusion matrix’s performance of the model in terms of accuracy, precision, recall, and f1-score.

Table 3. Model Evaluation Results Before and After Parameter Tunning

Model Evaluation Results Before Parameter Tunning				
Algoritma	Precision	Recall	F1-Score	Accuracy
SVM	99.22%	70.71%	82.58%	98.48%
Gradient Boosting	95.00%	62.98%	75.74%	97.95%
Model Evaluation Results After Parameter Tunning				
SVM	98.62%	79.00%	87.73%	98.88%
Gradient Boosting	95.93%	65.19%	77.63%	98.09%

The results of comparing the performance of SVM and gradient boosting models on the same dataset, based on the table above, show that the accuracy of the SVM model is superior to the use of the gradient boosting model. The accuracy of SVM and Gradient Boosting models also increases after parameter tuning. SVM is superior to gradient boosting by achieving the highest accuracy of 98.88%, while gradient boosting is only 98.09%. This research shows that constructing an SVM model that performs feature extraction with TF-IDF and performs parameter tuning can classify fake job postings very well.

Table 4 supports the findings of this research and compares the performance results of this study with those of previous research studies.

Table 4. Performance comparison with previous studies

Reference	Method	Accuracy
[33] 2020	Random Forest	98.27%
[13] 2021	XGB + Implementation of 2-step feature subset selection	97.94%
[19] 2022		XGBoost
[21] 2023	Bidirectional LSTM	98.71%
Proposed method	SVM + GridSearchCV	98.88%

Compared with previous research, this study excels in accuracy metrics, achieving a notable accuracy of 98.88%, surpassing the studies in the preceding years. Even when compared to the latest research, this study still outperforms the previous study by 0.17% in 2023.

#### 4. CONCLUSION

In this study, the classification of fake job postings was conducted using SVM and gradient-boosting algorithms. The research demonstrates the effectiveness of the SVM algorithm in classifying predictions of fake job postings. Feature extraction was performed using TF-IDF to identify keywords that distinguished between genuine and fake job postings. Parameter tuning was done to find a better combination of parameters to maximize accuracy. The results of the evaluation tests indicate success in improving the modeling accuracy of SVM and gradient-boosting algorithms after parameter tuning. The SVM algorithm proved to be the best model, achieving an accuracy of 98.88%, precision at 98.62%, recall at 7%, and an F1-Score of 87.73%. This study successfully increased accuracy by 0.17% compared to previous research. For future research, it is recommended to explore implementing data class balancing methods. Additionally, it is suggested that additional features be explored, ensemble algorithms be used, and cross-validation methods be applied to assist the model in generalizing to new data.

#### 5. ACKNOWLEDGEMENTS

We would like to thank the Artificial Intelligence and Data Mining Center (AIDMC) for its full moral and financial support. We also want to thank the MATRIK journal Editor for the opportunity to publish in the journal.

#### 6. DECLARATIONS

##### AUTHOR CONTRIBUTION

Rofik: Conceptualization, Methodology, Original draft. Roshan Aland Hakim: Original draft, Programming, Testing, Validation. Jumanto: Writing, Methodology, Editing, Supervision. Budi Prasetyo: Writing, Editing. Much Aziz Muslim: Writing, Editing, supervision.

##### FUNDING STATEMENT

This research was supported by the Artificial Intelligence and Data Mining Center (AIDMC), Universitas Negeri Semarang.

##### COMPETING INTEREST

The authors declare that they have no conflict of interest.

#### REFERENCES

- [1] S. Nematzadeh, F. Kiani, M. Torkamanian-Afshar, and N. Aydin, "Tuning hyperparameters of machine learning algorithms and deep neural networks using metaheuristics: A bioinformatics study on biomedical and biological cases," *Computational Biology and Chemistry*, vol. 97, p. 107619, Apr. 2022, <https://doi.org/10.1016/j.compbiolchem.2021.107619>. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S1476927121001894>
- [2] C. Vercellino, A. Scionti, G. Varavallo, P. Viviani, G. Vitali, and O. Terzo, "A Machine Learning Approach for an HPC Use Case: the Jobs Queuing Time Prediction," *Future Generation Computer Systems*, vol. 143, pp. 215–230, Jun. 2023, <https://doi.org/10.1016/j.future.2023.01.020>. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0167739X23000274>
- [3] A. F. Mulyana, W. Puspita, and J. Jumanto, "Increased accuracy in predicting student academic performance using random forest classifier," *Journal of Student Research Exploration*, vol. 1, no. 2, pp. 94–103, Jul. 2023, <https://doi.org/10.52465/josre.v1i2.169>. [Online]. Available: <https://shmpublisher.com/index.php/josre/article/view/169>

- [4] W. Lyu and J. Liu, "Artificial Intelligence and emerging digital technologies in the energy sector," *Applied Energy*, vol. 303, p. 117615, Dec. 2021, <https://doi.org/10.1016/j.apenergy.2021.117615>. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0306261921009843>
- [5] B. Chiraratanasopha and T. Chay-intr, "Detecting Fraud Job Recruitment Using Features Reflecting from Real-world Knowledge of Fraud," *Current Applied Science and Technology*, vol. 22, no. 6, Feb. 2022, <https://doi.org/10.55003/cast.2022.06.22.008>. [Online]. Available: <https://li01.tci-thaijo.org/index.php/cast/article/view/254033>
- [6] G. Othman Alandjani, "Online fake job advertisement recognition and classification using machine learning," *3C TIC: Cuadernos de desarrollo aplicados a las TIC*, vol. 11, no. 1, pp. 251–267, Jun. 2022, <https://doi.org/10.17993/3ctic.2022.111.251-267>. [Online]. Available: <https://www.3ciencias.com/articulos/articulo/online-fake-job-advertisement-recognition-and-classification-using-machine-learning/>
- [7] B. Alghamdi and F. Alharby, "An Intelligent Model for Online Recruitment Fraud Detection," *Journal of Information Security*, vol. 10, no. 03, pp. 155–176, 2019, <https://doi.org/10.4236/jis.2019.103009>. [Online]. Available: <http://www.scirp.org/journal/doi.aspx?DOI=10.4236/jis.2019.103009>
- [8] T. F. Waddell, H. Overton, and Robert McKeever, "Does sample source matter for theory? Testing model invariance with the influence of presumed influence model across Amazon Mechanical Turk and Qualtrics Panels," *Computers in Human Behavior*, vol. 137, p. 107416, Dec. 2022, <https://doi.org/10.1016/j.chb.2022.107416>. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0747563222002382>
- [9] K. Nanath and L. Olney, "An investigation of crowdsourcing methods in enhancing the machine learning approach for detecting online recruitment fraud," *International Journal of Information Management Data Insights*, vol. 3, no. 1, p. 100167, Apr. 2023, <https://doi.org/10.1016/j.ijime.2023.100167>. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S2667096823000149>
- [10] N. Jing, Z. Wu, S. Lyu, and V. Sugumaran, "Information credibility evaluation in online professional social network using tree augmented naive Bayes classifier," *Electronic Commerce Research*, vol. 21, no. 2, pp. 645–669, Jun. 2021, <https://doi.org/10.1007/s10660-019-09387-y>. [Online]. Available: <https://link.springer.com/10.1007/s10660-019-09387-y>
- [11] A. Amaar, W. Aljedaani, F. Rustam, S. Ullah, V. Rupapara, and S. Ludi, "Detection of Fake Job Postings by Utilizing Machine Learning and Natural Language Processing Approaches," *Neural Processing Letters*, vol. 54, no. 3, pp. 2219–2247, Jun. 2022, <https://doi.org/10.1007/s11063-021-10727-z>. [Online]. Available: <https://link.springer.com/10.1007/s11063-021-10727-z>
- [12] S. Dutta and S. K. Bandyopadhyay, "Fake Job Recruitment Detection Using Machine Learning Approach," *International Journal of Engineering Trends and Technology*, vol. 68, no. 4, pp. 48–53, Apr. 2020, <https://doi.org/10.14445/22315381/IJETT-V68I4P209S>. [Online]. Available: <https://ijettjournal.org/archive/ijett-v68i4p209s>
- [13] A. Mehboob and M. S. I. Malik, "Smart Fraud Detection Framework for Job Recruitments," *Arabian Journal for Science and Engineering*, vol. 46, no. 4, pp. 3067–3078, Apr. 2021, <https://doi.org/10.1007/s13369-020-04998-2>. [Online]. Available: <http://link.springer.com/10.1007/s13369-020-04998-2>
- [14] H. Sabita, F. Fitria, and R. Herwanto, "Analisa dan Prediksi Iklan Lowongan Kerja Palsu dengan Metode Natural Language Programming dan Machine Learning," *Jurnal Informatika*, vol. 21, no. 1, pp. 14–22, Jun. 2021, <https://doi.org/10.30873/ji.v21i1.2865>. [Online]. Available: <https://jurnal.darmajaya.ac.id/index.php/JurnalInformatika/article/view/2865>
- [15] S. Lal, R. Jiaswal, N. Sardana, A. Verma, A. Kaur, and R. Mourya, "ORFDetector: Ensemble Learning Based Online Recruitment Fraud Detection," in *2019 Twelfth International Conference on Contemporary Computing (IC3)*. Noida, India: IEEE, Aug. 2019, pp. 1–5, <https://doi.org/10.1109/IC3.2019.8844879>. [Online]. Available: <https://ieeexplore.ieee.org/document/8844879/>
- [16] J. Kim, H.-J. Kim, and H. Kim, "Fraud detection for job placement using hierarchical clusters-based deep neural networks," *Applied Intelligence*, vol. 49, no. 8, pp. 2842–2861, Aug. 2019, <https://doi.org/10.1007/s10489-019-01419-2>. [Online]. Available: <http://link.springer.com/10.1007/s10489-019-01419-2>

- [17] B. Baesens, S. Hppner, and T. Verdonck, "Data engineering for fraud detection," *Decision Support Systems*, vol. 150, p. 113492, Nov. 2021, <https://doi.org/10.1016/j.dss.2021.113492>. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0167923621000026>
- [18] M. Naud, K. J. Adebayo, and R. Nanda, "A machine learning approach to detecting fraudulent job types," *AI & SOCIETY*, vol. 38, no. 2, pp. 1013–1024, Apr. 2023, <https://doi.org/10.1007/s00146-022-01469-0>. [Online]. Available: <https://link.springer.com/10.1007/s00146-022-01469-0>
- [19] E. Baraneetharan, "Detection of Fake Job Advertisements using Machine Learning algorithms," *Journal of Artificial Intelligence and Capsule Networks*, vol. 4, no. 3, pp. 200–210, Oct. 2022, <https://doi.org/10.36548/jaicn.2022.3.006>. [Online]. Available: <https://irojournals.com/aicn/article/pdf/4/3/6>
- [20] M. Thanh Vo, A. H. Vo, T. Nguyen, R. Sharma, and T. Le, "Dealing with the Class Imbalance Problem in the Detection of Fake Job Descriptions," *Computers, Materials & Continua*, vol. 68, no. 1, pp. 521–535, 2021, <https://doi.org/10.32604/cmc.2021.015645>. [Online]. Available: <https://www.techscience.com/cmc/v68n1/41824>
- [21] A. S. Pillai, "Detecting Fake Job Postings Using Bidirectional LSTM," *International Research Journal of Modernization in Engineering Technology and Science*, Apr. 2023, <https://doi.org/10.56726/IRJMETS35202>. [Online]. Available: [https://www.irjmets.com/uploadedfiles/paper/issue\\_3\\_march\\_2023/35202/final/fin\\_irjmets1680354157.pdf](https://www.irjmets.com/uploadedfiles/paper/issue_3_march_2023/35202/final/fin_irjmets1680354157.pdf)
- [22] J. Jumanto, M. A. Muslim, Y. Dasril, and T. Mustaqim, "Accuracy of Malaysia Public Response to Economic Factors During the Covid-19 Pandemic Using Vader and Random Forest," *Journal of Information System Exploration and Research*, vol. 1, no. 1, pp. 49–70, Dec. 2022, <https://doi.org/10.52465/joiser.v1i1.104>. [Online]. Available: <https://shmpublisher.com/index.php/joiser/article/view/104>
- [23] S. Akuma, T. Lubem, and I. T. Adom, "Comparing Bag of Words and TF-IDF with different models for hate speech detection from live tweets," *International Journal of Information Technology*, vol. 14, no. 7, pp. 3629–3635, Dec. 2022, <https://doi.org/10.1007/s41870-022-01096-4>. [Online]. Available: <https://link.springer.com/10.1007/s41870-022-01096-4>
- [24] H. Tabassum, G. Ghosh, A. Atika, and A. Chakrabarty, "Detecting Online Recruitment Fraud Using Machine Learning," in *2021 9th International Conference on Information and Communication Technology (ICoICT)*. Yogyakarta, Indonesia: IEEE, Aug. 2021, pp. 472–477, <https://doi.org/10.1109/ICoICT52021.2021.9527477>. [Online]. Available: <https://ieeexplore.ieee.org/document/9527477>
- [25] M. R. Ningsih, K. A. H. Wibowo, A. U. Dullah, and J. Jumanto, "Global recession sentiment analysis utilizing VADER and ensemble learning method with word embedding," *Journal of Soft Computing Exploration*, vol. 4, no. 3, pp. 142–151, Sep. 2023, <https://doi.org/10.52465/josce.v4i3.193>. [Online]. Available: <https://shmpublisher.com/index.php/josce/article/view/193>
- [26] X. Wang, C. Wang, J. Yao, H. Fan, Q. Wang, Y. Ren, and Q. Gao, "Comparisons of deep learning and machine learning while using text mining methods to identify suicide attempts of patients with mood disorders," *Journal of Affective Disorders*, vol. 317, pp. 107–113, Nov. 2022, <https://doi.org/10.1016/j.jad.2022.08.054>. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0165032722009107>
- [27] P. Khandagale, A. Utekar, A. Dhonde, and P. S. S. Karve, "Fake Job Detection Using Machine Learning," *International Journal for Research in Applied Science and Engineering Technology*, vol. 10, no. 4, pp. 1822–1827, Apr. 2022, <https://doi.org/10.22214/ijraset.2022.41641>. [Online]. Available: <https://www.ijraset.com/best-journal/fake-job-detection-using-machine-learning>
- [28] G. Chaubey, P. R. Gavhane, D. Bisen, and S. K. Arjaria, "Customer purchasing behavior prediction using machine learning classification techniques," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 12, pp. 16133–16157, Dec. 2023, <https://doi.org/10.1007/s12652-022-03837-6>. [Online]. Available: <https://link.springer.com/10.1007/s12652-022-03837-6>
- [29] L. Zhou, H. Fujita, H. Ding, and R. Ma, "Credit risk modeling on data with two timestamps in peer-to-peer lending by gradient boosting," *Applied Soft Computing*, vol. 110, p. 107672, Oct. 2021, <https://doi.org/10.1016/j.asoc.2021.107672>. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S1568494621005937>

- [30] T. Wang, Y. Bian, Y. Zhang, and X. Hou, "Classification of earthquakes, explosions and mining-induced earthquakes based on XGBoost algorithm," *Computers & Geosciences*, vol. 170, p. 105242, Jan. 2023, <https://doi.org/10.1016/j.cageo.2022.105242>. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0098300422001911>
- [31] D. Valero-Carreras, J. Alcaraz, and M. Landete, "Comparing two SVM models through different metrics based on the confusion matrix," *Computers & Operations Research*, vol. 152, p. 106131, Apr. 2023, <https://doi.org/10.1016/j.cor.2022.106131>. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0305054822003616>
- [32] Z. M. Alhakeem, Y. M. Jebur, S. N. Henedy, H. Imran, L. F. A. Bernardo, and H. M. Hussein, "Prediction of Ecofriendly Concrete Compressive Strength Using Gradient Boosting Regression Tree Combined with GridSearchCV Hyperparameter-Optimization Techniques," *Materials*, vol. 15, no. 21, p. 7432, Oct. 2022, <https://doi.org/10.3390/ma15217432>. [Online]. Available: <https://www.mdpi.com/1996-1944/15/21/7432>