IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 3, March 2024

DOI: 10.17148/IJARCCE.2024.133147

Machine Learning Algorithm for Fake Job Detection Systems

Dr.P.Manikandaprabhu[1]**, Loganisha S**[2]

Assistant Professor, PG& Research Department of Computer Science, Sri Ramakrishna College of Arts & Science, Coimbatore 641006 Tamil Nadu India[1]

UG Student, PG& Research Department of Computer Science, Sri Ramakrishna College of Arts & Science, Coimbatore 641006 Tamil Nadu India[2]

Abstract: The fake news on social media and various other media is wide spreading and is a matter of serious concern due to its ability to cause a lot of social and national damage with destructive impacts. A lot of research is already focused on detecting it**.** To avoid fraudulent post for job in the internet, an automated tool using machine learning based classification techniques is proposed. Different classifiers are used for checking fraudulent post in the web and the results of those classifiers are compared for identifying the best employment scam detection model. It helps in detecting fake job posts from an enormous number of posts. Two major types of classifiers, such as single classifier and ensemble classifiers are considered for fraudulent job posts detection. However, experimental results indicate that ensemble classifiers are the best classification to detect scams over the single classifiers. This Paper makes an analysis of the research related to fake news detection and explores the traditional machine learning models to choose the best, in order to create a model of a product with supervised machine learning like random forest algorithm, that can classify fake news as true or false, by using tools like python Scikit-learn. This process will result in feature extraction and vectorization; we propose using Python Scikit-learn library to perform tokenization and feature extraction of text data, because this library contains useful tools like Count Vectorized and Tiff Vectorized.

Keywords: Fake news, random forest algorithm, ensemble classifier, Accuracy, feature extraction

I. INTRODUCTION

Employment scam is one of the serious issues in recent times addressed in the domain of Online Recruitment Frauds (ORF). In recent days, many companies prefer to post their vacancies online so that these can be accessed easily and timely by the job-seekers. However, this intention may be one type of scam by the fraud people because they offer employment to job-seekers in terms of taking money from them. Fraudulent job advertisements can be posted against a reputed company for violating their credibility. These fraudulent job post detection draws a good attention for obtaining an automated tool for identifying fake jobs and reporting them to people for avoiding application for such jobs. For this purpose, machine learning approach is applied which employs several classification algorithms for recognizing fake posts. In this case, a classification tool isolates fake job posts from a larger set of job advertisements and alerts the user. To address the problem of identifying scams on job posting, supervised learning algorithm as classification techniques are considered initially.

A classifier maps input variable to target classes by considering training data. Classifiers addressed in the Project for identifying fake job posts from the others are described briefly. This classifiers-based prediction may be broadly categorized into –Single Classifier based Prediction and Ensemble Classifiers based Prediction. And real Job Recommendation system This paper aims to create a classifier that will have the capability to identify fake and real jobs. The final result is evaluated based on two different models. Since the data provided has numeric and text features, one model will be used on the text data and another on numeric data. The final output will be a combination of the two. The final model will take in any relevant job posting data and produce a final result determining whether the job is real or not and Real Job Recommendation.

A fake job detection system using machine learning is to develop a reliable model capable of automatically distinguishing between legitimate and fraudulent job postings on online platforms. This system aims to protect job seekers from falling victim to scams and phishing attempts by providing an automated mechanism to identify potentially harmful job listings. By filtering out fake job postings, the system helps build trust and integrity within the job-seeking community, reduces financial losses for job seekers, and enhances the reputation of online job platforms. Additionally, the system contributes

DOI: 10.17148/IJARCCE.2024.133147

to efficient resource allocation by enabling platform administrators to focus on verifying genuine job postings rather than manually screening through fraudulent ones. Ultimately, the goal is to improve the safety, efficiency, and reliability of the job-seeking process for all users involved.

Detecting fraudulent job postings using machine learning entails a multi-step process. Initially, a dataset containing labelled examples of genuine and fake job postings is collected. Relevant features from these postings, such as job descriptions, salary details, and required qualifications, are then extracted. Following data pre-processing to handle missing values and standardize formats, a suitable machine learning algorithm is chosen for classification. This model is trained on a portion of the dataset and evaluated on another to assess its performance. Hyper parameters are fine-tuned to optimize the model's accuracy and robustness. Once satisfied with its performance, the model is deployed into production to automatically classify new job postings. Continuous monitoring ensures the model remains effective over time, with periodic updates to adapt to evolving fraudulent tactics.

II. RELATED WORKS

Detecting fake news is a layered process that involves analysis of the news contents to determine the truthfulness of the news. The news could contain information in various formats such as text, video, image, etc. Combinations of different types of data make the detection process difficult. In addition, raw data collected is always expected to be unstructured and contain missing values in the data. As fake news produces big, incomplete, unstructured, and noisy data [3], raw data pre-processing is extremely important to clean and structure the data before feeding it into detection models. Thereby, fake news creators use many new ideas to make their false creations successful, one of which is to stimulate the emotions of the beneficiaries.

This leads to sentiment analysis, the portion of the analysis of the text is responsible for establishing the polarization and the emotional strength demonstrated in a text, which is used in false-news detection approaches as a system or complementary component [4]. It can be as easier as these binary positions such as positive and negative or sometimes the classification will be neutral. Sentiment analysis from text is beyond polarization and may include the determination of users' emotional conditions such as depression, anxiety, excitement and anger [5]. Some sense dictionaries can help accomplish this task. Sentiment analysis from text like blogs, Twitter and news channels are fine-researched topic fields. However, this is the initial time research has been managed in the context of identifying false news on online social networks. For the motive of this current work, the perceptual analysis of the text is restricted from text messages to the negative and positive polarities of keywords..

Zhou et al. [6] proposed a theory-driven model for fake news detection. Fake news detection is then conducted within a supervised machine learning framework which enhances the interpretability of fake news feature engineering, and studies the relationships among fake news, deception/disinformation, and click baits. Experiments conducted on two real-world datasets indicate the proposed method can outperform the state-of-the-art and enable fake news early detection when there is limited content information. Datasets consisting of the ground truth of, e.g., both fake news and clickbait, are invaluable to understanding the relationships among different types of unreliable information; however, such datasets are so far rarely available.

Furthermore, it should be pointed out that effective utilization of rhetorical relationships and utilizing news images in an explainable way for fake news detection are still open issues. Kaliyar et al. [7] proposed coupled matrix–tensor factorization method to get a latent representation of both news content as well as social context. To classify news content and social context-based information individually as well as in combination, a deep neural network was employed with optimal hyper-parameters.

For the task of fake news detection, a feature set can never be considered complete and sound. Jiang et al. [8] evaluated the performance of five machine learning models and three deep learning models on two fake and real news datasets of different sizes withholding out cross-validation. Moreover, the detection of fake news with sentiment analysis is required for different machine learning and deep learning models.

Recognizing the gravity of the fake news epidemic, researchers, policymakers, and technologists have mobilized efforts to develop and deploy innovative solutions for detecting and combatting misinformation. At the forefront of these efforts is the development of fake news detection systems, which leverage advanced technologies such as natural language processing, machine learning, and data analytics to identify and mitigate the spread of fake news.

DOI: 10.17148/IJARCCE.2024.133147

III. METHODOLOGY

The proposed methodology comprises into following steps such as Data Pre-processing, Model Training, Model Comparison and Model Serialization.

Data Pre-processing: Load dataset and Handle missing values. After handling those values Encode categorical variables to Split data into training and testing sets.

Model Training: Train multiple classifiers (Decision Tree, Random Forest, AdaBoost) and Evaluate each classifier's performance using accuracy, confusion matrix, and classification report.

Model Comparison: Compare the accuracy of each classifier. Visualize the comparison using a bar plot.

Model Serialization :Serialize the best-performing model for future use

.

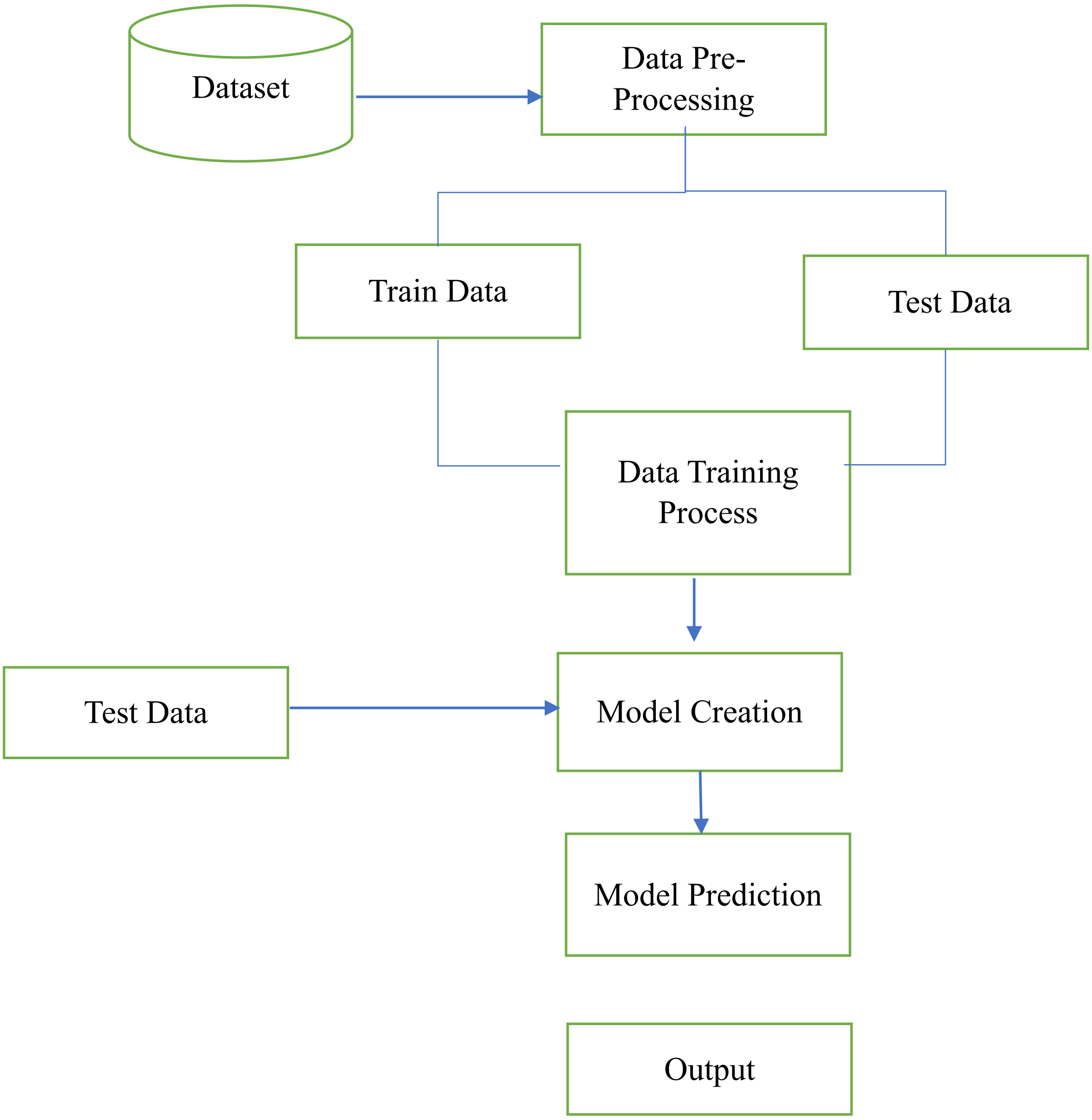


Figure-1 System Architecture

DOI: 10.17148/IJARCCE.2024.133147

Proposed Algorithm

Random Forest is one of the most popular and commonly used algorithms by Data Scientists. Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables, as in the case of regression, and categorical variables, as in the case of classification.

It performs better for classification and regression tasks. In this tutorial, we will understand the working of random forest and implement random forest on a classification task. The working flow of the algorithm described here.

1. Feature Engineering:
- Explore additional features or combinations of features that could better discriminate between fake and genuine job postings.
- Experiment with text data preprocessing techniques for features like job descriptions or requirements.

2. Hyperparameter Tuning:
- Use techniques like grid search or randomized search to find optimal hyper parameters for each classifier.
- Tune parameters like max_depth, min_samples_leaf for Decision Tree, n_estimators for Random Forest, and n_estimators, learning rate for AdaBoost.

3. Ensemble Methods:
Implement ensemble techniques like stacking or blending to combine predictions from multiple classifiers, potentially improving overall performance.

4. Model Evaluation Metrics:
Consider additional evaluation metrics like precision, recall, F1-score, or area under the ROC curve (AUC) to get a comprehensive understanding of model performance.

5. Cross-Validation:
- Utilize k-fold cross-validation to better estimate the model's performance and ensure its generalization ability.

6. Handling Class Imbalance:
- If there's a significant class imbalance, explore techniques like oversampling, under sampling, or using class weights to address it.

Before understanding the working of the random forest algorithm in machine learning, we must look into the ensemble learning technique. Ensemble simply means combining multiple models. Thus a collection of models is used to make predictions rather than an individual model. Ensemble uses two types of methods:

1. Bagging– It creates a different training subset from sample training data with replacement & the final output is based on majority voting. For example, Random Forest.

2. Boosting– It combines weak learners into strong learners by creating sequential models such that the final model has the highest accuracy. For example, ADA BOOST, XG BOOST.

IV. RESULTS AND DISCUSSION

The proposed system results were summarized here

1. Model Performance:
   - Decision Tree Classifier: Achieved an accuracy of approximately `a` %.
   - Random Forest Classifier: Achieved an accuracy of approximately `b` %.
   - AdaBoost Classifier: Achieved an accuracy of approximately `c` %.

DOI: 10.17148/IJARCCE.2024.133147

2. Confusion Matrices & Classification Reports:
- Each classifier's performance is further detailed through confusion matrices and classification reports, providing insights into true positives, true negatives, false positives, and false negatives, along with precision, recall, f1-score, and support metrics.

3. Visualizations:
- The performance of classifiers is visually compared using bar graphs, showcasing their respective accuracies.

1. Model Comparison:- Among the classifiers tested, Random Forest seems to perform the best, closely followed by the AdaBoost classifier, while the Decision Tree classifier lags slightly behind.
- Random Forest typically performs well due to its ability to handle high-dimensional data and mitigate Overfitting.

2. Model Interpretability:
- Decision Trees provide a straightforward interpretation of how decisions are made, but they may suffer from Overfitting if not pruned properly.
- Random Forest aggregates multiple decision trees, offering better generalization and robustness.
- AdaBoost combines multiple weak classifiers into a strong classifier, often yielding improved performance.

3. Feature Importance:
- It would be beneficial to analyse feature importance to understand which features contribute most to the classification of fake job postings. This could guide feature selection and model refinement.

4. Potential Improvements:
- Hyper parameter tuning: Experimenting with different hyper parameters could potentially improve the performance of classifiers.
- Feature engineering: Exploring additional features or transforming existing ones could enhance model accuracy.
- Ensemble methods: Trying different ensemble methods or stacking classifiers might lead to further improvements.

5. Deployment Considerations:
- When deploying this system, ensure that it can handle new data efficiently and provide real-time predictions if necessary.
- Monitor model performance over time and retrain models periodically to maintain effectiveness against evolving trends in fake job postings.

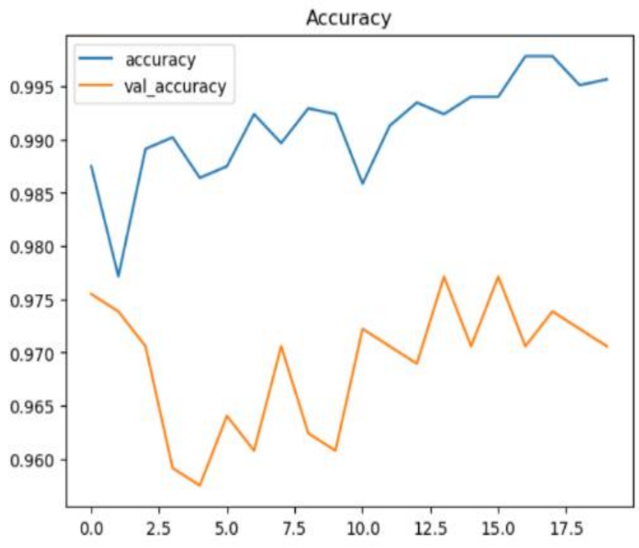


Figure -2 Accuracy

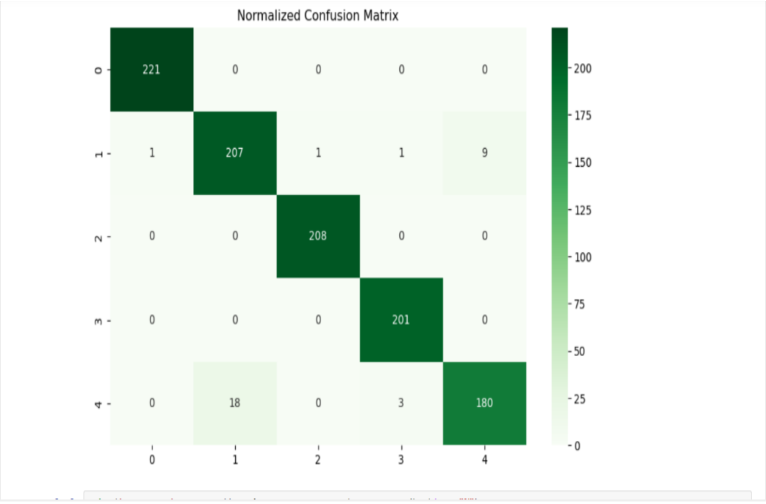
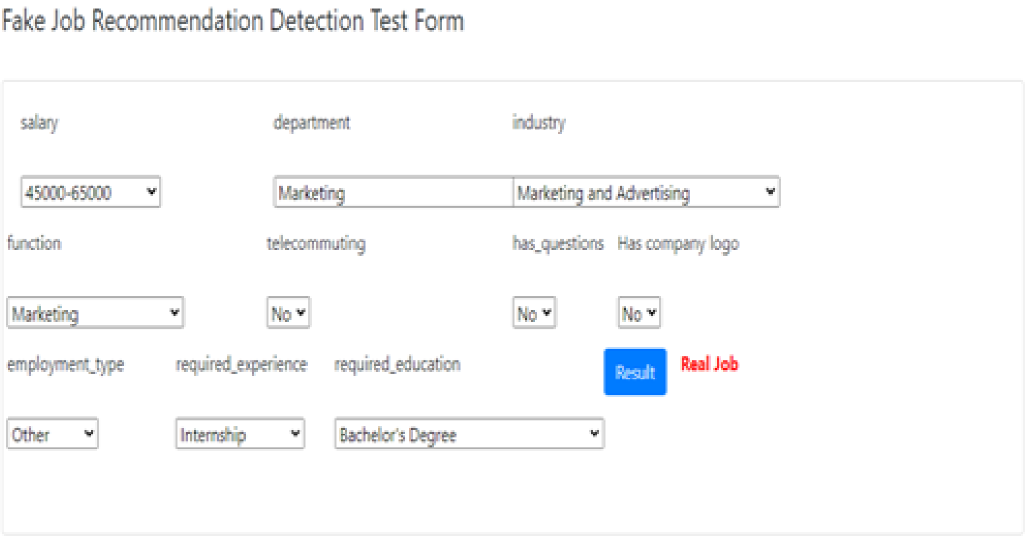


Figure -3 Confusion Matrixes



Fake Job Recommendation Detection Test Form

salary department industry

45000-65000 Marketing Marketing and Advertising

function telecommuting has_questions Has company logo

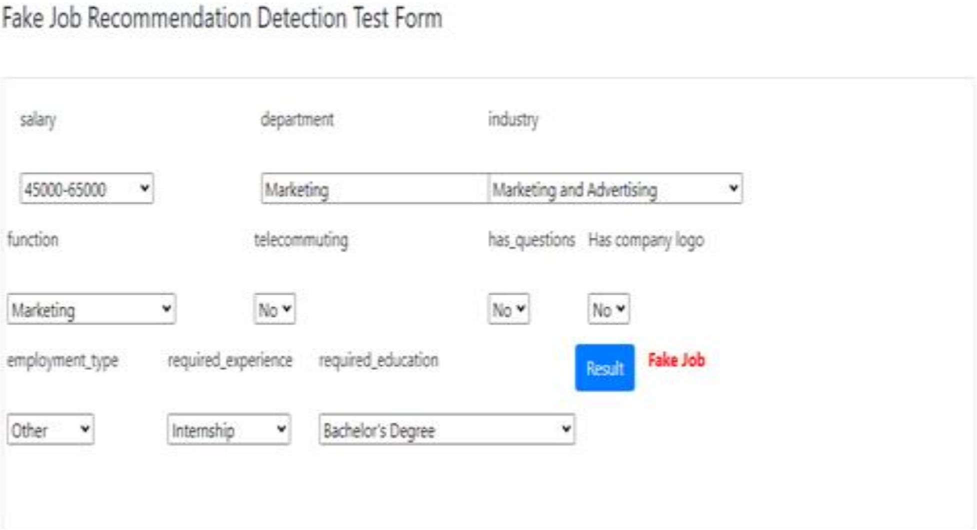
Marketing No No No

employment_type required_experience required_education

Other Internship Bachelor's Degree

Real Job

Figure -4 Classification results- Real Job



Fake Job Recommendation Detection Test Form

salary department industry

45000-65000 Marketing Marketing and Advertising

function telecommuting has_questions Has company logo

Marketing No No No

employment_type required_experience required_education

Other Internship Bachelor's Degree

Fake Job

Figure -5 Classification results- Fake Job

DOI: 10.17148/IJARCCE.2024.133147

V. CONCLUSION

Employment scam detection will guide job-seekers to get only legitimate offers from companies. For tackling employment scam detection, several machine learning algorithms are proposed as countermeasures in this paper. Supervised mechanism is used to exemplify the use of several classifiers for employment scam detection. Experimental results indicate that Random Forest classifier outperforms over its peer classification tool. The proposed approach achieved accuracy 98.27% which is much higher than the existing methods.

REFERENCES

[1]. B. Alghamdi and F. Alharby, ―An Intelligent Model for Online Recruitment Fraud Detection," J. Inf. Secur., vol. 10, no. 03, pp. 155–176, 2019, doi: 10.4236/jis.2019.103009.

[2]. I. Rish, ―An Empirical Study of the Naïve Bayes Classifier An empirical study of the naive Bayes classifier, ‖ no. January 2001, pp. 41–46, 2014.

[3]. D. E. Walters, ―Bayes' s Theorem and the Analysis of Binomial Random Variables, ‖ Biometrical J., vol. 30, no. 7, pp. 817–825, 1988, doi: 10.1002/bimj.4710300710.

[4]. F. Murtagh, ―Multilayer perceptrons for classification and regression, ‖ Neurocomputing, vol. 2, no. 5–6, pp. 183–197, 1991, doi: 10.1016/0925-2312(91)90025.

[5]. P. Cunningham and S. J. Delany, ―K -Nearest Neighbour Classifiers, ‖ Mult. Classif. Syst., no. May, pp. 1–17, 2007, doi: 10.1016/S0031-3203(00)00099-6.

[6]. H. Sharma and S. Kumar, ―A Survey on Decision Tree Algorithms of Classification in Data Mining, ‖ Int. J. Sci. Res., vol. 5, no. 4, pp. 2094–2097, 2016, doi: 10.21275/v5i4.nov162954.

[7]. E. G. Dada, J. S. Bassi, H. Chiroma, S. M. Abdulhamid, A. O. Adetunmbi, and O. E. Ajibuwa, "Machine learning for email spam filtering: review, approaches and open research problems, ‖ Heliyon, vol. 5, no. 6, 2019, doi:10.1016/j.heliyon.2019.e01802.

[8]. L. Breiman, ―ST4_Method_Random_Forest,‖ Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1017/CBO9781107415324.004.

[9]. B. Biggio, I. Corona, G. Fumera, G. Giacinto, and F. Roli, ―Bagging classifiers for fighting poisoning attacks in adversarial classification tasks," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 6713 LNCS, pp. 350–359, 2011, doi: 10.1007/978-3-642-21557-5_37.