Online Recruitment Fraud Detection: A Machine Learning-based Model for Ghanaian Job Websites

Delali Kwasi Dake
University of Education, Winneba
Department of Information and Communication Technology Education

ABSTRACT
The proliferation of online job websites has eased the difficulties in hiring and applying for jobs globally. Unfortunately, the risk of defrauding desperate job seekers exists with malicious recruiters taking advantage of the loopholes in the online recruitment process. The reactive approach to detecting online job fraud and the subsequent warnings on reputable job websites hasn't curtailed this spiteful act. The purpose of the study is to propose a machine learning model for proactive job fraud detection. In building the predictive model, a job fraud dataset from a job advertisement firm in Ghana was utilised. Using the 10-fold and the 5-fold cross-validation techniques, a job fraud detection model was built by comparing conventional and ensemble machine learning algorithms. The machine learning metrics, including accuracy, F1-score and the area under the curve (AUC) value, were reported and discussed. The findings show that the Random Forest traditional algorithm, with an accuracy of 91.86%, is best suited for the dataset. The investigation further indicates that information gain and chi-square feature selection mechanisms decreased classification accuracy marginally to 91.51%.

General Terms
Supervised Algorithm, Ensemble Learning, Job Fraud, Employment Scam, Online Recruitment Fraud

Keywords
Supervised Algorithm, Ensemble Learning, Job Fraud, Employment Scam, Online Recruitment Fraud

1. INTRODUCTION
Graduates' unemployment, especially in Africa, remains one of the global challenges confronting governmental agencies, policy organisations and academic institutions [1 – 3]. After graduation, each graduate desires a profession that will provide a good foundation for growth, prospects, and family. The disconnect, however, is that as the number of jobs available are few, students enrollment to tertiary institutions globally is on the rise [36]. According to UNESCO Institute of Statistics data, in June 2022, the total number of tertiary students worldwide doubled in the last two decades. South and West Asia, East Asia, and the Pacific had a staggering 240% increase in tertiary enrollment from 2000 to 2020, while in Central and Eastern Europe, enrollment increased by 84% during the same period. By 2020, Sub-Saharan African countries saw an average enrollment rise of 9.4%, with Mauritius having the highest gross tertiary enrollment of 40%. Within the same year, Ghana and Togo had an enrollment estimate of 15%, while Niger's estimate was 4.4%. The Organisation for Economic Co-operation and Development (OECD) 2021 report shows a relatively high unemployment rate across countries from 2017 to 2021 [37]. Among the OECD countries, South Africa recorded the highest tertiary unemployment figure of 13% in 2020, whiles the Czech Republic recorded the lowest figure of 1.4%. The OECD average unemployment figure stood at 4.3% for tertiary graduates aged 25 to 64 among member countries. The average unemployment figures increased to 6.4% for upper secondary, non-tertiary students within the same year. Another report by Statista 2022 shows that South Africa has the highest unemployment rate of 34%, followed by a 28% rate for Djibouti in Sub-Saharan Africa. Niger has the lowest unemployment figure of 1% whiles Ghana has a 5% figure. The average unemployment figure for Sub-Saharan Africa according to Statista 2022 stood at 8%. The frustrations of not getting a job, especially after tertiary education, deepen as graduates spend more time at home. The statistics indicate that more graduates will join the job hunt yearly, worsening an already fragile situation. As alluded to earlier, graduates, after successful education with acquired skills, expect to start a career with financial returns. The graduates search for jobs from diverse platforms, including social media, print media, companies and online job platforms [4, 5]. After several unsuccessful attempts, transitioning from graduation to employment leads to immense frustration with hasty decisions [6]. According to Indeed [38], the inability of a graduate to secure a job further leads to health-related problems, less family affection, unsavoury traits, and ultimately shrinks an economy. The ranking and reputation of tertiary institutions are equally affected when their graduates remain unemployed [39]. Therefore, the fierce competition in the job market requires academic institutions to prepare graduates adequately for the world of work [39, 1]. The preparation of learners should not be limited to 21st-century skills and knowledge acquisition but rather extend to curriculum vitae (CV) preparation and the prompt identification of fraudulent jobs. Internet penetration continues to rise globally. Statista [40] data reveals Denmark, the UAE, and Ireland as the leading countries with an internet penetration of 99%. The average global internet penetration stood at 63.1% in July 2022 [40]. In Africa, Morocco and Seychelles lead internet penetration at 84.1% and 79%, respectively. Ghana, where the study data was collected, has an internet penetration of 53%. The popularity of job search platforms has root linkages to the growth in internet penetration worldwide. An online job website has the standard functionality of linking employers to job seekers [41] with other relevant advantages, including [42]

a) Time-saving. Easy job application anywhere and anytime

b) Search filters. Diverse customised filters to define job limits and standards

c) User-friendly. Primary, modern job websites have good navigation and are easy to use.

d) Economical. Aside from internet costs, hard-copy printing and travelling expenses are limited

e) Networking. Job seekers have the opportunity to network with employers and job seekers

Even though the advantages of using job platforms are numerous, recruitment fraud remains an issue from employers and the negligence of job platforms [42, 41]. Reputable job websites have several mechanisms to track fraudulent employers and job posts but largely remain unsuccessful due to the desperation of job seekers and the diverse strategies the fraudulent recruiter adopts [38, 43].

Machine learning (ML) has recently been at the forefront of research with immense applications in agriculture, healthcare, education, industry, transportation and social media. In agriculture, ML has been used mainly in yield prediction, disease detection, weed detection, species recognition, livestock management, soil management, crop quality and water management [7, 8]. In healthcare, ML has been used significantly in wearable devices to detect patients' health, develop new drugs, medical imaging, personalised medicine, disease outbreak prediction and diagnostic chatbots [9, 10]. In education, ML has been utilised primarily in learner sentiment modelling, learner groupings, academic performance prediction, learner attrition prediction, students behaviour modelling and online learning analytics [11 – 13]. In industry, ML has seen advancement in predictive maintenance, equipment automation, logistics and inventory management, robotics and e-commerce [14 – 16]. In transportation, ML algorithms has seen integration in route prediction, traffic prediction, environmental condition monitoring, surveillance and security [17, 18]. Social media has seen diverse applications of ML for market segmentation, friend recommendation, spam filtering, advertisement and sentiment analysis [19, 20].

Available data shows that graduates' unemployment figures remain relatively high, especially in African countries [37], with adverse implications for graduates, academic institutions, and the global economy [3]. In contrast, internet penetration universally has been ascending [40], laying a foundation for the proliferation of online job websites. Furthermore, the number of mobile devices in 2022 increased from 14.91 billion in 2021 to 15.96 billion, with an estimated projection of 18.22 billion by 2025 [40]. The ease of applying for jobs, internet penetration and the rise in mobile devices have made online job platforms attractive for both recruiters and job seekers. The employer has a primary objective of hiring the best candidates, which is possible when a higher number of candidates apply for the advertised position. Since there is a disparity between the number of advertised jobs and the number of candidates applying, the desperation to get employed becomes high [21, 22]. The job seeker is exposed to employment risks and mostly becomes victims of financial extortions from malicious recruiters who lower ethics for financial gains.

1.1 Purpose of Study

In line with the problem definition, online job filtering to detect high-risk job posts from the recruiter has become mandatory. This study proposes a machine-learning model for Ghanaian job websites by comparing conventional and ensemble algorithms using the 10-fold and the 5-fold cross-validation techniques. The following research questions guided the study:

(1) What is the optimum accuracy between conventional and ensemble algorithms using the 10-fold and 5-fold cross-validation techniques?

(2) What are the highest values of the conventional and ensemble algorithms' F-measure and AUC-ROC after classification?

(3) To what extent has the accuracy, F-measure, and AUC-ROC values changed after implementing the information gain and chi-square feature selection mechanisms?

2. REVIEW OF LITERATURE

The fraud perpetrated by recruiters using online job websites has seen the inscription of various warning statements and policy documents on the websites of reputable job platforms [38, 44, 45]. Aside from demanding payment before job seekers get the ghost job, spiteful recruiters sexually harass female applicants and utilise chronic means to get vital information from candidates [46].

The first aspect of the review discusses the use of natural language processing (NLP) for text-based job classification. Text-based job classification includes the identification of keywords in job descriptions, application methods and other relevant job modules. With text-based job classification, the labelling of a text without verification or complaint from a job seeker can negatively affect the labelling of a job post as fraudulent.

Naudé et al., [23] implemented empirical rule set-based features, transformer models, word embeddings and bag-of-word (BoW) feature selection mechanism on Employment Scam Aegean Dataset (EMSCAD) to categorise the text-based job types as corporate identity theft, real job, multi-level marketing and identity theft. The identity theft categorisations are based on demands for non-CV personal information from candidates and referrals to an external website to complete the application process. Corporate identity theft consists of illicit job advertisements pretending to originate from legitimate firms, whereas multi-level marketing refers to commission-based schemes that induce candidates to send the fraudulent job advertisements to other candidates. The three feature selection schemes were tested with AdaBoost (AB), Gradient Boosting (GB), Random Forest (RF), Support Vector Machine (SVM), Decision Tree (CART), k-Nearest Neighbor (KNN), Stochastic Gradient Descent (SGD) and Logistic Regression (LR) algorithms for F1-score and Matthews correlation coefficient (MCC). The results show that the GB with bag-of-word feature selection has the best F1-score of 0.88.

Tabassum et al., [24] tracked 4000 instances of datasets from different job websites in Bangladesh using the EMSCAD reference template. The aggregated data contained 3699 real jobs and 301 fake jobs. During data pre-processing for classification, the Term Frequency Inverse Document Frequency (TF-IDF) was utilised to determine the weight of frequency occurring text in the dataset. The LR, CART, RF, AB, GB, Voting Classifier (VC) and LightGBM (LG) classifiers were compared for accuracy, and the result was referenced with the EMSCAD dataset. The Voting Classifier algorithm emerged as the best classifier, with an accuracy of 95.34% after feature selection.

Dutta & Bandyopadhyay [25] utilised the EMSCAD dataset and implemented categorical encoding of some selected features to create a vector of words. The Naïve Bayes (NB), Decision Tree (DT), Multi-Layer Perceptron (MLP), AB, GB and Random Tree (RT) classification algorithms were used to build the classification model. The accuracy, F1-score, Cohen-kappa and Mean Squared Error (MSE) performance metrics were compared during the 80% to 20% training and testing

ratio. RT emerged as the best classifier with an accuracy of 98.27% and a Cohen-kappa score of 0.74.

Amaar et al., [26] proposed an ML-based model for job fraud detection using BoW and TF-IDF feature selection mechanisms on EMSCAD dataset consisting of 17,014 and 866 legitimate and fraudulent jobs, respectively. The dataset imbalances and over-fitting problems were solved using the adaptive synthetic (ADASYN) oversampling technique. The RF, LR, Support Vector Machine (SVM), Extra Trees Classifier (ETC), and MLP algorithms were implemented to develop the classification model. The study also compared long short term memory (LSTM), convolutional neural networks (CNN) and gated recurrent unit (GRU) deep learning algorithms with the supervised learning algorithms. The ETC model achieved the highest accuracy of 99.9% using TF-IDF feature selection and ADASYN oversampling technique.

Mahbub et al., [27] implemented machine learning algorithms with localised data from Australia for online recruitment fraud detection. Local data was sorted from the Gumtree website in Australia with 2,276 job instances. The J48 DT, NB, RF and JRip rule-based algorithms were implemented with the feature space technique in WEKA to determine the accuracy and F1-score of the classifiers. RF has the highest accuracy of 91.86% using content-based and contextual features.

Khandagale et al., [47] built a fake job detection model using NB, SVM, RF and LR algorithms on EMSCAD dataset from Kaggle. The TF-IDF feature selection technique was utilised during data pre-processing to process the string attributes and measure essential terms. Simulation results show that, the RF has the highest accuracy of 97%

The second aspect of the review uses only classification devoid of natural language processing (NLP) during data pre-processing. The classification method without text analysis is relevant when the attributes are correctly labelled based on the information retrieved from the job description and the application method.

Alghamdi & Alharby [28] utilised the EMSCAD dataset and implemented the ensemble RF classifier after initial feature selection using SVM. Since the EMSCAD dataset was labelled, the pre-processing stage avoided using NLP-based feature selection techniques in the string attribute types. The 5-fold and 10-fold cross-validation techniques were applied to the dataset, and the accuracy was examined. The RF ensemble performed with an accuracy of 97.41% using the 10-fold cross-validation technique.

Mehboob & Malik [29] removed unnecessary information from the EMSCAD dataset with a two-step strategy to select the best subset of attributes. The 17,880 imbalanced instances of data was reduced to 470 legitimate ads and 470 fraudulent ads. The information gain, gain ratio and correlation coefficient feature selection techniques were implemented to select the top 18-features for classification. The NB, KNN, DT, MLP, SVM, RF and XGBoost (XGB) algorithms were implemented using the 10-fold cross-validation technique to build the classifier. XGB, from the classification results, has the highest accuracy of 97.94%

2.1 Findings and Gaps in Literature Review

The literature review shows the use of EMSCAD dataset from Kaggle as the standard dataset for online recruitment fraud detection. Since most researchers could not find a country-based suitable dataset to tailor recruitment fraud detection on job websites to respective countries, improving accuracy from the EMSCAD dataset became a priority. Aside from Tabassum et al., [24] and Mahbub et al., [27], who utilised datasets outside EMSCAD, the rest of the limited literature resorted to machine learning performance metric comparisons. The second observation from the review is the lack of comparison between cross-validation techniques. Alghamdi & Alharby [28] is the only study that compared the 10-fold and the 5-fold cross-validation techniques but failed to compare the findings with other algorithms.

In this proposed research, recruitment data was sorted from a local jobs website in Ghana to understand the strategies of the malicious recruiters. A machine learning model is then proposed using conventional and ensemble algorithms with the 10-fold and 5-fold cross-validation techniques. The relevant performance metrics, including accuracy, F1-score and AUC-ROC score, are discussed before and after applying feature selection methods.

3. RESEARCH METHODOLOGY

The research methodology, as shown in Figure 1, begins with the Jobweb Ghana job scam dataset, cleaned during data pre-processing. The dataset is checked for data imbalances using the SMOTE over-sampling technique. The conventional and ensemble ML algorithms are compared using the 10-fold and the 5-fold cross-validation techniques in building the model. Feature selection is then implemented to check performance against the initial accuracy, and the best ML model is implemented

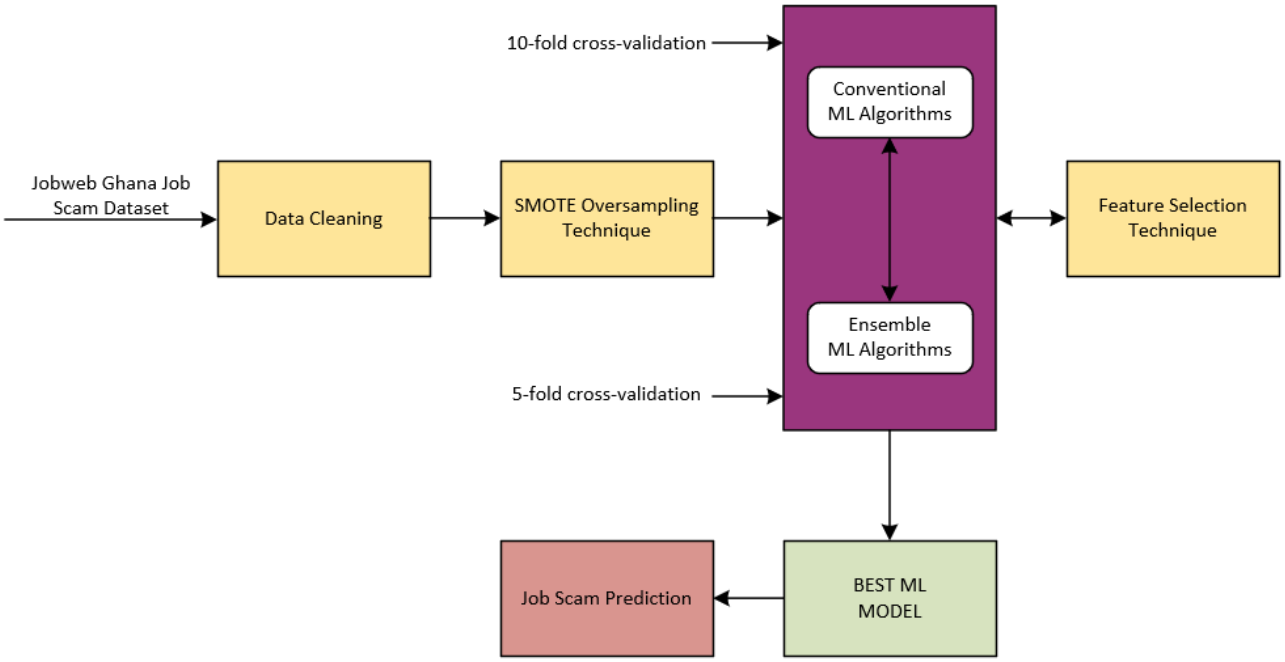

Figure 1: Methodological Framework

3.1 The Dataset

The statistics on employment fraud was extracted from Jobweb Ghana (www.jobwebghana.com), the second-largest job board in Ghana [48]. Tracking fraudulent jobs based on complaints from job seekers and verification from Jobweb Ghana was done between 2019 and 2021. Since the researcher is the founder of Jobweb Ghana, the tracking of malicious job posts from employers was well documented. The job fraud dataset consists of 499 instances, with 27.86% fraudulent and 72.14% genuine jobs. As shown in Table 1, recruiters mainly use eleven attributes when posting a job advert on Jobweb Ghana. Under the company name, 71.74% of employers reserve the actual name of their firms when placing job adverts. Approximately 60.32% of the job descriptions are long, with 33.47% disclosing salary ranges. A whopping 82.57% of recruiters do not use their company logo but mostly disclose the job application closing date. The data also reveals that 64.33% of recruiters match the job description to standard qualifications but mostly do not disclose their telephone numbers. Employers prefer email, constituting 86.77% as a means for candidates to send applications, with most job opportunities located in cities, especially Accra and Kumasi

Table 1: Attributes for Job Scam Dataset

Attribute	**Options**
Company Name	Open; Reserved
Job Description	Long; Moderate; Short
Salary Disclosure	Yes; No
Multiple Categories	Yes; No
Employment Type	Full-time; Part-time; Contract
Application Deadline	Disclosed; Undisclosed
Company Logo	Yes; No
Degree Required	Standard; Low
Location	City; Town
Company Phone Number	Yes; No
Main Method of Application	Email; Address; Website
Class	Genuine; Fraudulent

3.2 SMOTE Oversampling Filter

The SMOTE filter was applied to the minority class to increase data instances by 150% and prevent over-fitting. As shown in Figure 2, the fraudulent default class instances of 139 has increased to 347 with the application of the SMOTE over-sampling filter to deal with data imbalances.

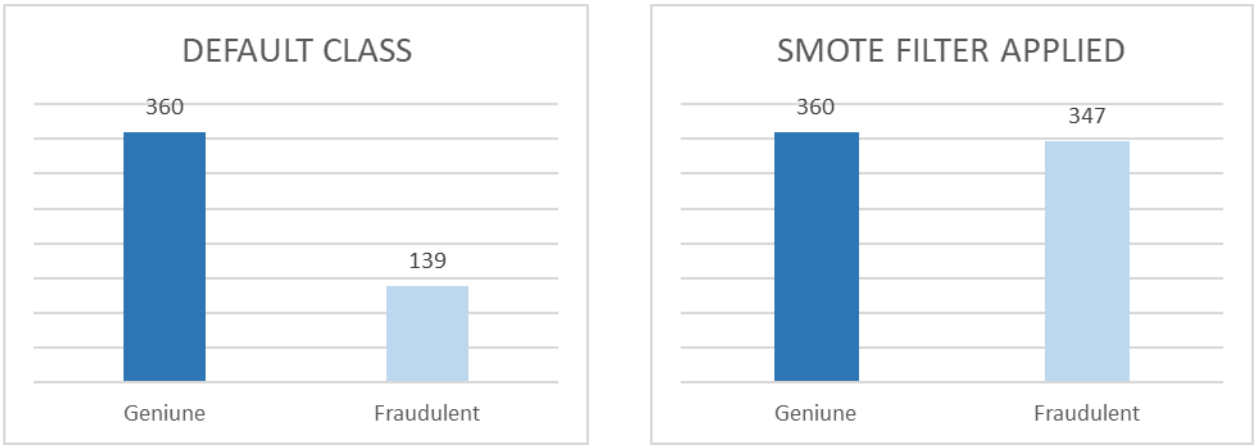

Figure 2: (a) Default class label; (b) SMOTE filter

3.3 Conventional and Ensemble ML Algorithms

Ensemble learning, a multi-classifier system, incorporates multiple machine learning models in the prediction process. In ensemble learning, weak learners are combined to serve as base models to achieve higher classification results [30].

AdaBoost (AB) is an ensemble algorithm based on the boosting technique. As an iterative algorithm, AB uses several weak learners on the same training dataset to construct a high-performing final algorithm. AdaBoost uses two weights, one for each weak learning algorithm and the other for the training set sample [31].

The Bagging algorithm forms multiple bootstrapped subsamples from a dataset to form a single classifier. Bagging creates a diverse predictive model using different uniform samples from the learning algorithm. Bagging is based on bootstrap aggregation [31]

Conventional machine learning algorithms depend on labelled training data to predict class labels. The efficient training of data in supervised learning generates a model for solving classification and regression-related problems [32].

The Support Vector Machine (SVM) maps inputs to high-dimensional feature space to perform non-linear classification. SVM draws a margin between classes intending to reduce the classification error by increasing the distance between the margin and the classes [32].

The Naïve Bayes (NB) is a probabilistic machine learning algorithm based on the Bayes theorem. NB uses independent features to calculate a posterior probability with the assumption of an event occurring based on another event termed conditional probability [33].

The Decision Tree (DT) uses simple decision rules initiated from the root to the terminal node, the leaf. The DT algorithm uses splitting to form sub-nodes hierarchically until all possible outcomes are realised at the final terminal [34].

The Random Forest (RF) is a supervised learning algorithm constructed from DT. The RF algorithm is based on ensemble techniques and uses the mean of different trees to predict class outcomes successfully. Increasing the number of trees in RF enhances the accuracy of the model generated [35].

4. RESULTS AND ANALYSIS

This section compares conventional and ensemble machine learning results using the 10-fold and the 5-fold cross-validation techniques in WEKA. The dataset utilised is the one generated after implementing the SMOTE over-sampling filter.

4.1 10-fold and 5-fold Cross-Validation Experiments

The 10-fold cross-validation technique iteratively uses nine parts of the dataset, initially divided into ten for training, with the remaining one for testing. The process is randomly repeated ten times and helps prevent over-fitting. Instead of the ten parts, the 5-fold cross-validation uses five divisions with four parts for training and the remaining one for testing iteratively.

Research Question 1: What is the optimum accuracy between conventional and ensemble algorithms using the 10-fold and 5-fold cross-validation techniques?

As depicted in Tables 2 and 3, among the conventional algorithms, RF using the 10-fold technique has the highest accuracy of 91.94% compared to 91.51% of RF using the 5-fold. The best ensemble classifiers have the same accuracy of 91.80% for the Bagging (RF) using the 10-fold and AdaBoost (RF) using the 5-fold technique. Compared to the ensemble approach, the traditional RF algorithm with 10-fold cross-validation has the highest accuracy at 91.94%, a gap of 0.14%.

Table 2: ML Performance metrics for the 10-fold technique

Classifier	Accuracy (%)	F1-score	ROC Value
SVM	90.09	0.901	0.901
NB	89.95	0.900	0.958
RF	**91.94**	**0.919**	**0.962**
J48 DT	91.09	0.911	0.941
AdaBoost (SVM)	90.95	0.909	0.959
AdaBoost (NB)	89.96	0.900	0.923
AdaBoost (RF)	91.23	0.912	0.946
AdaBoost (DT)	91.09	0.911	0.961
Bagging (SVM)	90.81	0.908	0.931
Bagging (NB)	90.24	0.902	0.958
Bagging (RF)	**91.80**	**0.918**	**0.967**

Bagging (DT)	91.51	0.915	0.964

Table 3: ML Performance metrics for the 5-fold technique

Classifier	Accuracy (%)	F1-score	ROC Value
SVM	90.10	0.901	0.901
NB	90.10	0.901	0.958
RF	**91.51**	**0.915**	**0.963**
J48 DT	90.66	0.907	0.938
AdaBoost (SVM)	89.82	0.898	0.951
AdaBoost (NB)	90.10	0.901	0.919
AdaBoost (RF)	**91.80**	**0.918**	0.944
AdaBoost (DT)	91.65	0.917	0.960
Bagging (SVM)	89.96	0.900	0.930
Bagging (NB)	90.10	0.901	0.958
Bagging (RF)	91.65	0.917	**0.963**
Bagging (DT)	90.95	0.909	0.960

Research Question 2: What are the highest values of the conventional and ensemble algorithms' F-measure and AUC-ROC after classification?

The F1-score and the AUC-ROC are two essential metrics aside from accuracy, which does not explain the confusion matrix. Whiles the F-measure is the harmonic mean of precision and recall, the AUC-ROC determines the extent of separability between the classes in the confusion matrix. In F1-score and AUC-ROC, the algorithm with the highest value to 1 is more significant in precision, recall and class seperability measures. Tables 2 and 3 show that the conventional RF from the 10-fold technique still has the highest F1-score of 0.919. However, the ensemble Bagging (RF) using the 10-fold technique has the highest AUC-ROC value of 0.967.

Research Question 3: To what extent has the accuracy, F-measure, and AUC-ROC values changed after implementing the information gain and chi-square feature selection mechanisms?

In response to research question 3, the information gain and chi-square feature selection were implemented to determine the ranking of attributes with respect to classification accuracy. As illustrated in Table 4, the impact of the attributes are listed in descending order. The worst-performing features, including location, employment type, company phone number and company logo, were removed before the second classification

Table 4: Feature Selection Mechanism

Attributes	Information Gain	Chi-Square
Degree Required	0.424	373.02
Multiple Categories	0.230	213.92
Salary Disclosure	0.210	195.58
Application Deadline	0.175	164.15
Company Name	0.115	102.08
Job Description	0.093	89.47
Main Method of Application	0.087	65.92
Company Logo	0.069	60.63
Company Phone Number	0.021	9.77
Employment Type	0.009	9.06
Location	0.0007	0.740

The 10-fold cross-validation outperformed the 5-fold cross after applying feature selection, as shown in Tables 5 and 6. The conventional RF algorithm still has the highest accuracy of 91.51% and an F1-score of 0.915. The accuracy, however, decreased compared to the classification results with no feature selection technique in Table 2. The Bagging (RF) using the 10-fold cross-validation has the AUC-ROC value reduced to 0.966 after feature selection. In summary, the classification accuracy after applying feature selection decreased from 91.94% to 91.54% with the conventional RF algorithm.

Table 5: 10-fold cross-validation - after feature selection

Classifier	Accuracy (%)	F1-score	ROC Value
SVM	89.53	0.895	0.896
NB	89.53	0.895	0.958
RF	**91.51**	**0.915**	**0.962**
J48 DT	88.82	0.888	0.939
AdaBoost (SVM)	89.25	0.892	0.957
AdaBoost (NB)	89.53	0.895	0.920
AdaBoost (RF)	91.51	0.915	0.952
AdaBoost (DT)	90.66	0.907	0.965
Bagging (SVM)	89.95	0.900	0.935
Bagging (NB)	89.39	0.894	0.959
Bagging (RF)	**91.23**	**0.912**	**0.966**
Bagging (DT)	88.68	0.887	0.961

Table 6: 5-fold cross-validation - after feature selection

Classifier	Accuracy (%)	F1-score	ROC Value
SVM	89.82	0.898	0.898
NB	89.53	0.895	0.959
RF	90.66	0.907	0.962
J48 DT	89.53	0.895	0.935
AdaBoost (SVM)	89.11	0.891	0.954
AdaBoost (NB)	89.53	0.895	0.912
AdaBoost (RF)	**91.24**	**0.912**	0.947
AdaBoost (DT)	90.81	0.908	0.958
Bagging (SVM)	89.53	0.895	0.937
Bagging (NB)	89.53	0.895	0.959
Bagging (RF)	91.09	0.911	**0.963**
Bagging (DT)	88.96	0.890	0.959

5. DISCUSSION AND FINDINGS

The increasing internet penetration globally has made online job posting and recruitment attractive for employers. However, fake employers have taken advantage of desperate job seekers via various means, including enticing salary scales, reducing the qualification standard, selecting multiple categories and hiding company information to defraud. Most studies from the review relied on EMSCAD dataset from Kaggle to develop a predictive model for online recruitment fraud detection. Dutta & Bandyopadhyay [25] achieve an accuracy of 98.27% on EMSCAD dataset using the Random Tree algorithm without cross-validation. Amaar et al. [26], also without cross-validation, has an accuracy of 99.9% with the Extra Tree Classifier on the EMSCAD dataset. Alghamdi & Alharby [28] has an accuracy of 97.41% with the Random Forst algorithm but did not implement an over-sampling technique to check class imbalances. Mehboob & Malik [29] utilised the EMSCAD dataset, with the XGBoost algorithm having the highest accuracy of 97.94%. In their study, class imbalances was manually performed. Tabassum et al. [24] utilised Bangladesh job websites with the highest accuracy of 95.34% using the Voting Classifier algorithm. They, however, did not use any over-sampling and cross-validation techniques. Mahbub et al. [27], using the Gumtree data, achieved an accuracy of 91.86% with the Random Forest algorithm but did not implement any cross-validation technique.

In this proposed solution for online job fraud detection in Ghana, the cross-validation technique and the SMOTE over-sampling filter to prevent class imbalances were utilised. With the 10-fold cross-validation technique, the Random Forest emerged as the best classifier with an accuracy of 91.94%. The result is similar to Mahbub et al. [27] job fraud detection model using Gumtree data from Australia. In their study, the RF also emerged as the best classifier. Even though Tabassum et al. [24] used Bangladesh job website data, their accuracy was high but misleading since the cross-validation technique that checks over-fitting was not implemented.

6. CONCLUSION AND FUTURE WORK

Graduates' unemployment and desperation in getting a job has compromised them as victims of recruitment fraud globally and in Ghana. Even with recruitment fraud notices from reputable online job websites, some graduates still fall prey to these malicious fraudsters. Most job fraud victims pay and sometimes compromise their private details to the disguised recruiters. In Ghana, online job fraud is surging to the extent that reputable companies put occasional notices in print media to warn job seekers. Since the researcher is the founder of Jobweb Ghana, enough data was taken over the period about the main attributes of fraudulent recruiters and their cunning job descriptions.

The proposed solution compared conventional and ensemble machine learning algorithms using the 10-fold and the 5-fold cross-validation techniques. In dealing with the minority class instances, the SMOTE over-sampling filter was used to increase class imbalances by 150%. The SMOTE filter prevents over-fitting and a classification bias to the majority class. Aside from the classification accuracy, the F1-score, which measures the mean between precision and recall, was analysed. The AUC-ROC value, which signals the level of separability between the class labels, was also reported. The 10-fold cross-validation technique performed better than the 5-fold cross-validation even after feature selection. After feature selection, the classification accuracy of the conventional and ensemble machine learning algorithms decreased marginally. The feature selection mechanism, in conclusion, was not relevant to the dataset.

One aspect of future research is to compare conventional machine algorithms and ensemble learning schemes to deep neural networks. Deep neural networks with multiple hidden layers can increase the machine learning output metrics. Another valuable study is implementing Reinforcement Learning (RL) software agents to detect online recruitment fraud by continuously analysing malicious job posts.

7. REFERENCES

[1] M. Osmani, V. Weerakkody, N. Hindi, and T. Eldabi, "Graduates employability skills: A review of literature against market demand," *J. Educ. Bus.*, vol. 94, no. 7, pp. 423–432, 2019, doi: 10.1080/08832323.2018.1545629.

[2] O. Fakunle and H. Higson, "Interrogating theoretical and empirical approaches to employability in different global regions," *High. Educ. Q.*, vol. 75, no. 4, pp. 525–534, 2021, doi: 10.1111/hequ.12345.

[3] L. Graham, L. Williams, and C. Chisoro, "Barriers to the labour market for unemployed graduates in South Africa," *J. Educ. Work*, vol. 32, no. 4, pp. 360–376, 2019, doi: 10.1080/13639080.2019.1620924.

[4] R. C. Sewell, M. Martin, S. Barnett, and C. Jenter, "Graduating to success in employment: ow social media can aid college students in the job search," *John J. Heldrich Cent. Work. Dev.*, no. September, pp. 1–12, 2011, [Online]. Available: http://staging.helc01-002.svmsolutions.com/sites/default/files/content/Graduating_to_Success_Brief.pdf.

[5] S. Monteiro, L. Almeida, and A. García-Aracil, "'It's a very different world': work transition and employability of higher education graduates," *High. Educ. Ski. Work. Learn.*, vol. 11, no. 1, pp. 164–181, 2021, doi: 10.1108/HESWBL-10-2019-0141.

[6] C. S. Johnston, "A Systematic Review of the Career

Adaptability Literature and Future Outlook," *J. Career Assess.*, vol. 26, no. 1, pp. 3–30, 2018, doi: 10.1177/1069072716679921.

[7] K. G. Liakos, P. Busato, D. Moshou, S. Pearson, and D. Bochtis, "Machine learning in agriculture: A review," *Sensors (Switzerland)*, vol. 18, no. 8, pp. 1–29, 2018, doi: 10.3390/s18082674.

[8] A. Sharma, A. Jain, P. Gupta, and V. Chowdary, "Machine Learning Applications for Precision Agriculture: A Comprehensive Review," *IEEE Access*, vol. 9, pp. 4843–4873, 2021, doi: 10.1109/ACCESS.2020.3048415.

[9] H. K. Bharadwaj *et al.*, "A Review on the Role of Machine Learning in Enabling IoT Based Healthcare Applications," *IEEE Access*, vol. 9, pp. 38859–38890, 2021, doi: 10.1109/ACCESS.2021.3059858.

[10] P. Doupe, J. Faghmous, and S. Basu, "Machine Learning for Health Services Researchers," *Value Heal.*, vol. 22, no. 7, pp. 808–815, 2019, doi: 10.1016/j.jval.2019.02.012.

[11] D. K. Dake and E. Gyimah, "Using sentiment analysis to evaluate qualitative students' responses," *Educ. Inf. Technol.*, 2022, doi: 10.1007/s10639-022-11349-1.

[12] C. Buabeng-andoh, "Using Machine Learning Techniques to Predict Seasonal Rainfall-New," vol. 2022, 2022.

[13] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 10, no. 3, pp. 1–21, 2020, doi: 10.1002/widm.1355.

[14] T. Wuest, D. Weimer, C. Irgens, and K. D. Thoben, "Machine learning in manufacturing: Advantages, challenges, and applications," *Prod. Manuf. Res.*, vol. 4, no. 1, pp. 23–45, 2016, doi: 10.1080/21693277.2016.1192517.

[15] T. T. Ademujimi, M. P. Brundage, and V. V. Prabhu, "A Review of Current Machine Learning Techniques Used in Manufacturing Diagnosis," *IFIP Adv. Inf. Commun. Technol.*, vol. 513, pp. 407–415, 2017, doi: 10.1007/978-3-319-66923-6_48.

[16] Z. Ge, Z. Song, S. X. Ding, and B. Huang, "Data Mining and Analytics in the Process Industry: The Role of Machine Learning," *IEEE Access*, vol. 5, pp. 20590–20616, 2017, doi: 10.1109/ACCESS.2017.2756872.

[17] F. Zantalis, G. Koulouras, S. Karabetsos, and D. Kandris, "A review of machine learning and IoT in smart transportation," *Futur. Internet*, vol. 11, no. 4, pp. 1–23, 2019, doi: 10.3390/FI11040094.

[18] Y. Sirisha, S. S. Garlapati, D. Dubey, G. Shashank Reddy, and N. Shashi Kiran, "Traffic Prediction for Intelligent Transportation System," *Ymer*, vol. 21, no. 5, pp. 499–504, 2022, doi: 10.37896/YMER21.05/56.

[19] R. V. Belfin, E. Grace Mary Kanaga, and S. Kundu, "Application of Machine Learning in the Social Network," *Recent Adv. Hybrid Metaheuristics Data Clust.*, no. June, pp. 61–83, 2020, doi: 10.1002/9781119551621.ch4.

[20] B. T.k., C. S. R. Annavarapu, and A. Bablani, "Machine learning algorithms for social media analysis: A survey," *Comput. Sci. Rev.*, vol. 40, p. 100395, 2021, doi: 10.1016/j.cosrev.2021.100395.

[21] T. Petry, C. Treisch, and M. Peters, "Designing job ads to stimulate the decision to apply: a discrete choice experiment with business students," *Int. J. Hum. Resour. Manag.*, vol. 33, no. 15, pp. 3019–3055, 2022, doi: 10.1080/09585192.2021.1891112.

[22] J. A. Rios, G. Ling, R. Pugh, D. Becker, and A. Bacall, "Identifying Critical 21st-Century Skills for Workplace Success: A Content Analysis of Job Advertisements," *Educ. Res.*, vol. 49, no. 2, pp. 80–89, 2020, doi: 10.3102/0013189X19890600.

[23] M. Naudé, K. J. Adebayo, and R. Nanda, "A machine learning approach to detecting fraudulent job types," *AI Soc.*, no. 2016, 2022, doi: 10.1007/s00146-022-01469-0.

[24] H. Tabassum, G. Ghosh, A. Atika, and A. Chakrabarty, "Detecting Online Recruitment Fraud Using Machine Learning," *2021 9th Int. Conf. Inf. Commun. Technol. ICoICT 2021*, pp. 472–477, 2021, doi: 10.1109/ICoICT52021.2021.9527477.

[25] S. Dutta and S. K. Bandyopadhyay, "Fake job recruitment detection using machine learning approach," *Int. J. Eng. Trends Technol.*, vol. 68, no. 4, pp. 48–53, 2020, doi: 10.14445/22315381/IJETT-V68I4P209S.

[26] 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 Process. Lett.*, vol. 54, no. 3, pp. 2219–2247, 2022, doi: 10.1007/s11063-021-10727-z.

[27] S. Mahbub, E. Pardede, and A. S. M. Kayes, "Online Recruitment Fraud Detection: A Study on Contextual Features in Australian Job Industries," *IEEE Access*, vol. 10, no. May, pp. 82776–82787, 2022, doi: 10.1109/ACCESS.2022.3197225.

[28] 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.

[29] A. Mehboob and M. S. I. Malik, "Smart Fraud Detection Framework for Job Recruitments," *Arab. J. Sci. Eng.*, vol. 46, no. 4, pp. 3067–3078, 2021, doi: 10.1007/s13369-020-04998-2.

[30] Z. Asghari Varzaneh, M. Shanbehzadeh, and H. Kazemi-Arpanahi, "Prediction of successful aging using ensemble machine learning algorithms," *BMC Med. Inform. Decis. Mak.*, vol. 22, no. 1, p. 258, 2022, doi: 10.1186/s12911-022-02001-6.

[31] X. Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, "A survey on ensemble learning," *Front. Comput. Sci.*, vol. 14, no. 2, pp. 241–258, 2020, doi: 10.1007/s11704-019-8208-z.

[32] A. Dey, "Machine Learning Algorithms: A Review," *Int. J. Comput. Sci. Inf. Technol.*, vol. 7, no. 3, pp. 1174–1179, 2016, [Online]. Available: www.ijcsit.com.

[33] F. J. Yang, "An implementation of naive bayes classifier," *Proc. - 2018 Int. Conf. Comput. Sci. Comput. Intell. CSCI 2018*, pp. 301–306, 2018, doi: 10.1109/CSCI46756.2018.00065.

[34] B. Charbuty and A. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," *J. Appl. Sci. Technol. Trends*, vol. 2, no. 01, pp. 20–28, 2021, doi:

10.38094/jastt20165.

[35] Z. Wu, W. Lin, Z. Zhang, A. Wen, and L. Lin, "An Ensemble Random Forest Algorithm for Insurance Big Data Analysis," *Proc. - 2017 IEEE Int. Conf. Comput. Sci. Eng. IEEE/IFIP Int. Conf. Embed. Ubiquitous Comput. CSE EUC 2017*, vol. 1, pp. 531–536, 2017, doi: 10.1109/CSE-EUC.2017.99.

[36] UNESCO (2020, December 16). UNESCO IESALC report reveals that access to higher education increased from 19% to 38% in the last two decades. https://www.iesalc.unesco.org/en/2020/12/16/unesco-iesalc-report-reveals-that-access-to-higher-education-increased-from-19-to-38-in-the-last-two-decades/

[37] OECD (2021). Unemployment rates by education level. https://data.oecd.org/unemp/unemployment-rates-by-education-level.htm

[38] Indeed (2022, May 6). How Unemployment Affects Individuals and the Economy. https://www.indeed.com/career-advice/career-development/effects-unemployment

[39] QS (2022). QS Graduate Employability Rankings 2022. https://www.topuniversities.com/university-rankings/employability-rankings/2022

[40] Statista (2022, August 3). Unemployment rate in Africa as of 2022, by country. https://www.statista.com/statistics/1286939/unemployment-rate-in-africa-by-country

[41] Workable (2022). What is a job board?. https://resources.workable.com/hr-terms/what-is-a-job-board

[42] Siôn Phillpott (2021, March 31). The Advantages and Disadvantages of Online Recruitment. https://www.careeraddict.com/advantages-and-disadvantages-of-online-recruitment

[43] Glassdoor (2022, August 18). Fraudulent job postings. https://help.glassdoor.com/s/article/Fraudulent-job-postings?language=en_US

[44] Clark (2019, October 13). How to Spot and Avoid Online Job Scams. https://www.linkedin.com/pulse/how-spot-avoid-online-job-scams-biron-clark/

[45] JobwebGhana (2021, August 13). Terms & Conditions. https://jobwebghana.com/terms-conditions/

[46] Radloff (2020, January 28). Recruitment Fraud: How to Protect Your Brand. https://www.nasrecruitment.com/2020/01/28/recruitment-fraud-how-to-protect-your-brand/

[47] Khandagale, P., Utekar, A., Dhonde, A., & Karve, S. (2022). Fake job detection using machine learning. *International Journal for Research in Applied Science & Engineering Technology (IJRASET), 10(5),* 1822–1826. https://doi.org/10.22214/ijraset.2022.41641

[48] Golearners Hub (2019). Top job websites in Ghana. https://golearnershub.com/list-of-five-best-job-search-portals-ghana/