Knowledge Discovery in Research Security Practices among Scientists using Machine Learning Techniques (A Case Study of Faculty of Science, University of Ibadan)

O. Osunade
University of Ibadan, Nigeria
Department of Computer Science

I.T. Ayorinde
University of Ibadan, Nigeria
Department of Computer Science

B.I. Ayinla
University of Ibadan, Nigeria
Department of Computer Science

ABSTRACT
The burgeoning digitization of scientific research and the concurrent proliferation of sensitive data emphasize the pressing need to investigate and enhance research security practices among scientists. This study outlines a comprehensive knowledge discovery endeavor that leverages machine learning techniques to analyze a survey designed to uncover insights into the current state of research security practices within the scientific community. The study focuses on a survey conducted among scientists in the Faculty of Science, University of Ibadan to gain insights into their awareness, adoption, and perceptions of research security measures. It seeks to identify the prevailing trends, challenges, and gaps in security practices that may compromise the integrity and confidentiality of scientific research. Through analysis, machine learning and visualization techniques, the study uncovered valuable patterns and correlations within the survey data. The knowledge discovery process in this study involved examining factors such as researchers' status, years of experience, knowledge of dual-use research, medium of data storage, collaboration experience, training on research security and risk identification among others. The outcomes of this research encompass the identification of common security vulnerabilities, best practices, and potential areas for improvement in safeguarding scientific research data. Hence, the results of this study is a potential tool to inform policy development, enhance security awareness initiatives, and guide the scientific community in strengthening its defenses against threats to research integrity.

General Terms
Knowledge Discovery with Machine Learning Techniques.

Keywords
Research security, Dual-use technology, Machine learning, Best practices, Bootstrapping.

1. INTRODUCTION
Research security simply means safeguarding research enterprise against the misappropriation of research and development to the detriment of national or economic security. It also safeguards research integrity, and foreign government interference [1,2]. Some of the areas to safeguard in research are Intellectual Property theft, which deals with the stealing of another person's idea, invention or creative expression and Forced Technology Transfers which deal with the strategic state acquisition of foreign technologies to enhance domestic capacity while simultaneously reducing the benefits to innovators [2].

Research Security can also be referred to as the ability to identify possible risks to research work through unwanted access, interference, or theft and the measures that minimize these risks and protect the inputs, processes, and products that are part of scientific research and discovery [3].

Research security involves the actions that protect research communities from actors and behaviours that pose economic, strategic, and/or national and international security risks. It is an emerging area for many researchers, institutions, and governments [4].

Inability to safeguard research can lead to diminished trust and confidence in the research data and results, loss of research data, loss of exclusive control over intellectual property, patent opportunities and potential revenue, legal or administrative consequences, loss of potential future partnerships and of course, tarnished reputation [3].

Machine Learning (ML) is a branch of Artificial Intelligence (AI) that brings out the power of data in new ways. It helps computer systems learn and improve from experience by developing computer programs that can automatically access data and perform tasks via predictions and detections [5]. It has been known from both literature and past works that machine learning and data mining techniques thrive better with big data. Discovering knowledge comes with the ability of a ML algorithm to infer new knowledge from a new set of data after training it with some data [6]. According to [7], it is necessary to transform the raw data into clear and practical information to make predictions. But in cases where few data are being used, the desired result may not be achieved. Hence, resulting into a technique called bootstrapping.

Data boosting, which is also known as data augmentation or bootstrapping, is a technique used in machine learning and data science to improve the performance of models by increasing the amount and diversity of training data. It involves generating additional training examples by applying various transformations or perturbations to the existing data. This process can help address issues related to overfitting, generalization, and class imbalance.

Hence, this study carries out a survey of research security practices among scientists in the Faculty of Science, University of Ibadan and the result has shown that few researchers are aware of what research security is while majority are not even aware of dual-use research. The bootstrapping technique employed revealed that formal training on research security is

of utmost importance to researchers.

Some of the research questions answered are:

1. Do researchers know if their work has dual-use?
2. What are the research security practices of researchers?
3. What has been the foreign research collaboration experience of scientists?

2. RELATED WORKS

According to [8], academic fraud, which is also lack of research security is a rising threat. Schemes to defraud funding bodies, institutions and researchers for personal gain are not a modern invention within academia but one that threatens to topple the integrity of research practice. These manifest in the form of internal research misconduct and external predatory practice, the former perpetrated by the over-ambitious and the latter by organizations predating on unsuspecting researchers. Such academic fraud can undermine academic integrity, profoundly influence key legislation, and cause societal damage. Hence, the author called for a major reform of the academic system in order to overcome these difficulties. He further divided the measures used into detection and prevention methods. Detection methods include peer-review, replication, whistle blowing, external review bodies and digital solutions among others while Prevention methods include awareness, data repositories, institutional and editorial policies, punishment and deterrence, transparency indices, and changes to the 'publish or perish' mentality. These solutions are as of yet immature, flawed or in need of major revision but do have some potential in overcoming the rising threat of academic fraud.

According to [9], collaborative research usually comes with ethical issues which threatens research security. While research collaboration can be a productive way to advancing research skills, it also comes with some potential risks such as miscommunication, conflict, plagiarism, data misuse and ethical violations among others [9,10]. Collaborative research must be started with a clear goal, expectations and roles. Collaborators credentials, reputation and previous works must be checked at the beginning . Collaborators must establish regular and transparent communication channels like email, phone, video call, or online platforms. Since data is the core of any research project, it must be treated with care and respect.

In collaborative research, one of the most important aspects of research ethics and integrity is to acknowledge and credit the collaborators for their contributions. There must be an agreement on the authorship order, roles, and responsibilities of all collaborators. Collaborator's works must be cited and referenced appropriately while any form of plagiarism, duplication or fabrication should be avoided. Also, any potential conflicts of interest or biases that may affect collaboration or results must be disclosed [10].

According to [4], best practices in research are usually being underpinned by research security and integrity. While research security deals with protecting the processes and outputs of research, research integrity deals with the adherence to professional values, principles, and best practices which uphold the validity, social relevance, responsibility and quality of research. All these form the base on which researchers can collaborate in a fair, innovative, open and trusted research environment. Research integrity ensures that individuals can be confident in the advancement of research knowledge and in the dissemination of its results.

Some best practices that can be adapted in securing research involve being current with cybersecurity practices, system authentication and security, data backup, data encryption, installation of anti-virus software and firewall and protection of research labs among others [11].

The author in [7] examined the efficacy of a novel Index Mapped Ordinal Encoding Method (IMOEM) for machine learning algorithm in terms of precision, recall and accuracy. The performance of the IMOEM built on crime detection dataset with respect to precision, recall, f-score and accuracy results were significantly effective. The model performed exceptionally well with no loss of accuracy either in precision or recall values, especially when applied to the decision tree based Models. Researchers are therefore encouraged to embrace the use of IMOEM for machine learning algorithm as it helps discover new patterns.

The authors in [12] developed a deep learning model (which is also a machine learning technique) that accurately classify edible and poisonous mushrooms using multi-layer perceptron (MLP) neural network. The MLP with principal component analysis (PCA) was found to perform better than the one without PCA. Hence, the ability to better differentiate between edible and poisonous mushrooms using the PCA model will save more lives. This study also shows the efficacy of ML techniques in discovering new knowledge.

3. METHODOLOGY

Two different methods were employed in analyzing the data collected for this study. Google Data Studio was used to analyze the raw data while Bootstrapping and machine learning techniques were used to learn and discover new knowledge.

3.1 Analysis of the Raw Data

This study used the descriptive survey approach to answer the research questions. A survey of 27 questions was administered to 201 academic and research staff of the Faculty of Science, University of Ibadan, Nigeria, out of which only 47 responded. The survey was divided into six sections. Section 1 was the demographic data of respondents. Section 2 had 7 questions and focused on research experiences, section 3 had a question on research protection mechanism, section 4 had 6 questions on research purpose, section 5 had 6 questions focused on computing skill, and section 6 had 4 questions about risk identification in research. The survey was administered using Microsoft Forms for a 1-week period from 17 September to 24 September, 2023. The authors had to make repeated appeals for the survey to be filled. The raw data collected was analyzed using simple statistical methods such as frequency. Google Data Studio was used for the data analysis. The result is discussed in section 4.1.

3.2 Bootstrapping and Machine Learning Techniques

In addition to the statistical analysis, the data was boosted and run on three machine learning algorithms. Bootstrapping is a resampling technique that was employed to improve the model performance due to the small dataset. Bootstrapping is a machine learning method that helps estimate the uncertainty of a statistical model. The original 47 dataset sampling were involved in the replacement and generating of multiple new datasets of the same size as the original. Each of these new datasets is then used to calculate the desired statistic, such as the mean or standard deviation. This process is repeated

multiple times, and the resulting values are used to construct a probability distribution for the desired statistic. This technique was used to estimate the accuracy of the models, validate its performance, and identify areas that need improvement.

Due to the imbalance nature of the dataset used in this research. The Two most popular over-sampling techniques were adopted to build reliable and generalizable models. The Synthetic Minority Oversampling Technique (SMOTE) synthesizes new minority instances between existing minority instances, as illustrated in Figure 8. It randomly picks up the minority class and calculates the K-nearest neighbor for that particular point. Finally, the synthetic points are added between the neighbors and the chosen spot. The synthetic dataset was used to build the two models, which are Extreme Gradient Boosting (XGBoosting) and Traditional Random Forest (RF), to perform prediction or classifications of the unknown dataset being the most effective ensemble algorithms.

The second oversampling method was Random Oversampler that oversamples the minority class data as done by SMOTE oversampler. The Random Oversample model picks random data points from the existing datasets and generates a group of synthetic datasets. The dataset was used to equally build two ensemble models to enhance feature engineering and validate the performance of the models.

4. IMPLEMENTATION AND DISCUSSION OF RESULTS

The analysis of the results for both raw and boosted data are discussed in this section. While descriptive statistics was used to analyze the result for the raw data, machine learning techniques were used for the ensemble models.

4.1 Discussion of the Data Analysis Results

Google Data Studio, which is also known as Looker was used to analyse the raw data used in this study. The results are presented in this section.

4.1.1 Demographics

A total of 47 respondents completed the survey out of a population of 201. There are more male than female respondents as shown in Table 1. The number of respondents represent 23% of the population of researchers in the Faculty of Science, University of Ibadan.

Table 1: Gender of Respondents

Male	25
Female	22
Total	47

4.1.2 Academic / Research Status

The survey showed that the questionnaire was completed mainly by Professors (26%) and Senior Lecturers (26%). There is a low number of early career researchers (Assistant Lecturer to Lecturer 1) who responded as shown in Figure 1.

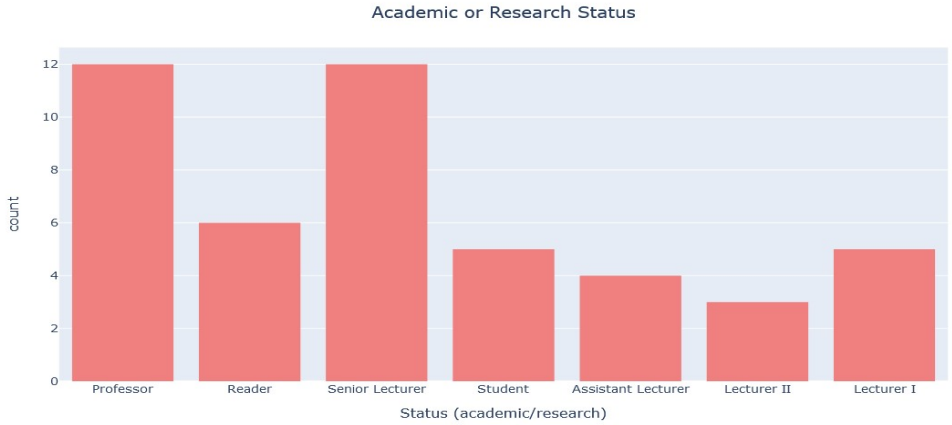

Figure 1: Status (Academic/Research)

4.1.3 Result of Respondents According to Discipline

The computing research area had more respondents than other fields of research as shown in Figure 2. The mathematical or physical research area had the second highest number of respondents. This may be because of the importance attached to data for research by computing and mathematical researchers. The survey was however completed by researchers in all science research areas.

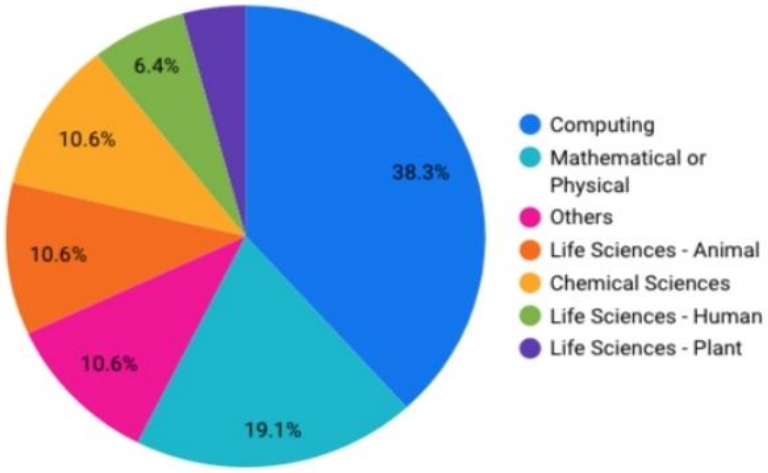

Figure 2: Classification of Respondents According to Discipline

4.1.4 Length of Research Experience

In Figure 3, researchers with 11-15 years research experience make 26% of the respondents followed by 17% of respondents with 6-10 years of research experience. There is a wide range of research experience available in the Faculty of Science, University of Ibadan.

Figure 3: Length of Research Experience (In Years)

4.1.5 The Research Security Practices of Researchers

Figure 4 answers the Research Question 2: "What are the research security practices of researchers?". The agreement, copyright and creative commons license are the three mostly used protection mechanism by respondents as highlighted in Figure 4. The use of agreement as a protection mechanism has been used by all levels of researchers in the Faculty of Science, University of Ibadan. The other protection mechanisms such as Patent have been minimally used by all categories of staff.

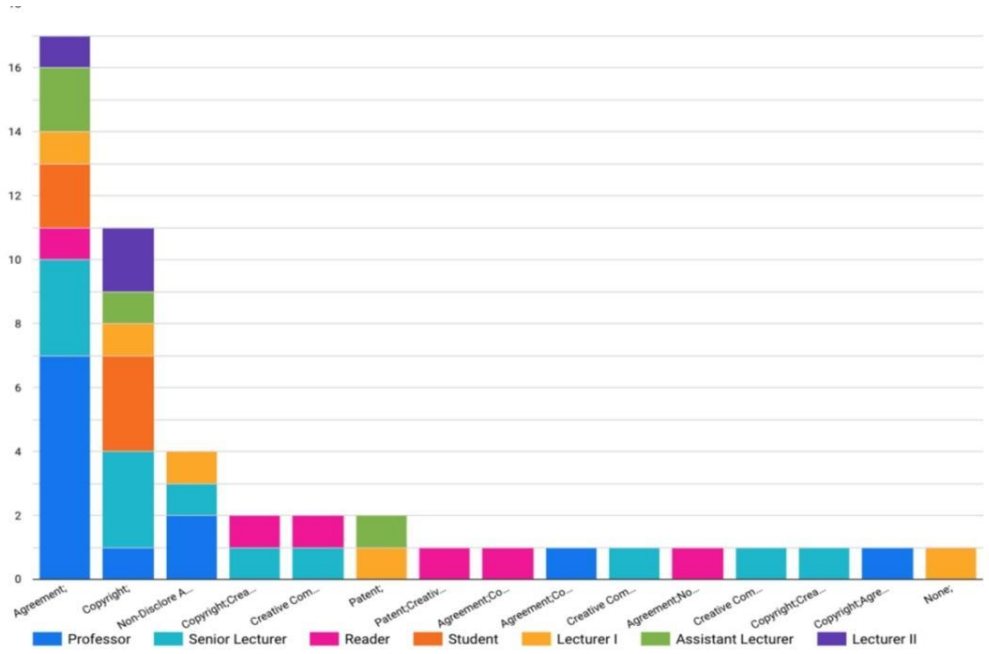

Figure 4: Knowledge Protection Mechanism Used by Respondents

4.1.6 Medium of Data Storage

Figure 5 depicts how research data is stored amongst the respondents. The storage of research data is mostly (42.6%) done using local storage devices such as hard disks and flash drives. The CD-ROM is no longer in use since most computing devices do not have them anymore.

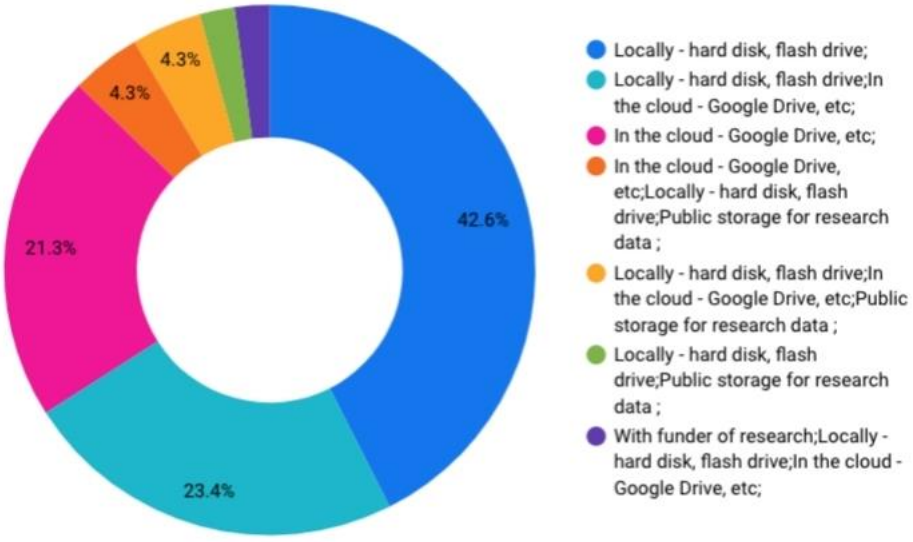

Figure 5: Mechanism of Data Storage by Respondents

4.1.7 Means of Accessing Internet by Researchers

The mobile phone is the mostly used device (55.3%), as shown in Figure 6, for Internet access by respondents. The versatility of the device and relationship with the device owner have been responsible for this.

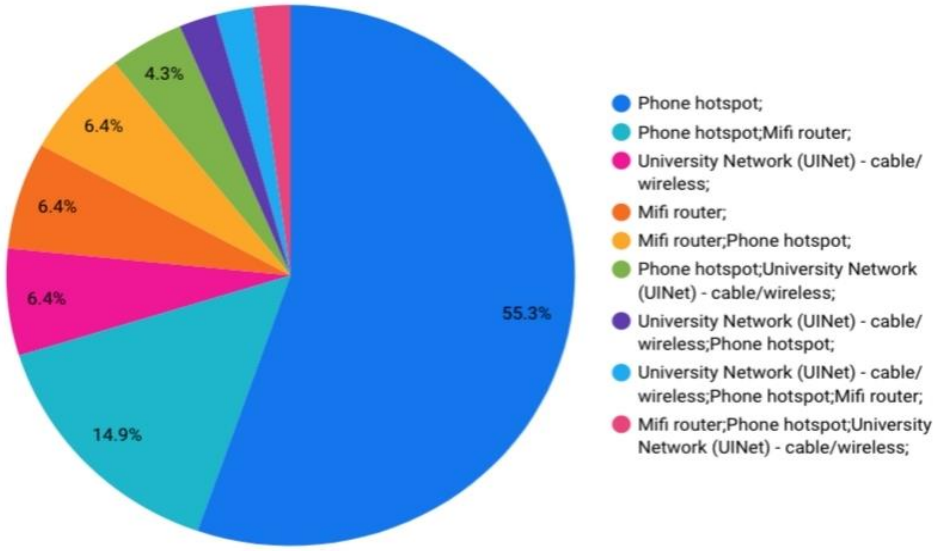

Figure 6: Means of Accessing Internet by Respondents

4.1.8 Risks Resulting From Stolen Data

The top three identified risks for researchers if their research data is stolen are career slowdown (20.69 %), litigation (13.8%) and Promotion delay (10.34%) as indicated by Figure 7.

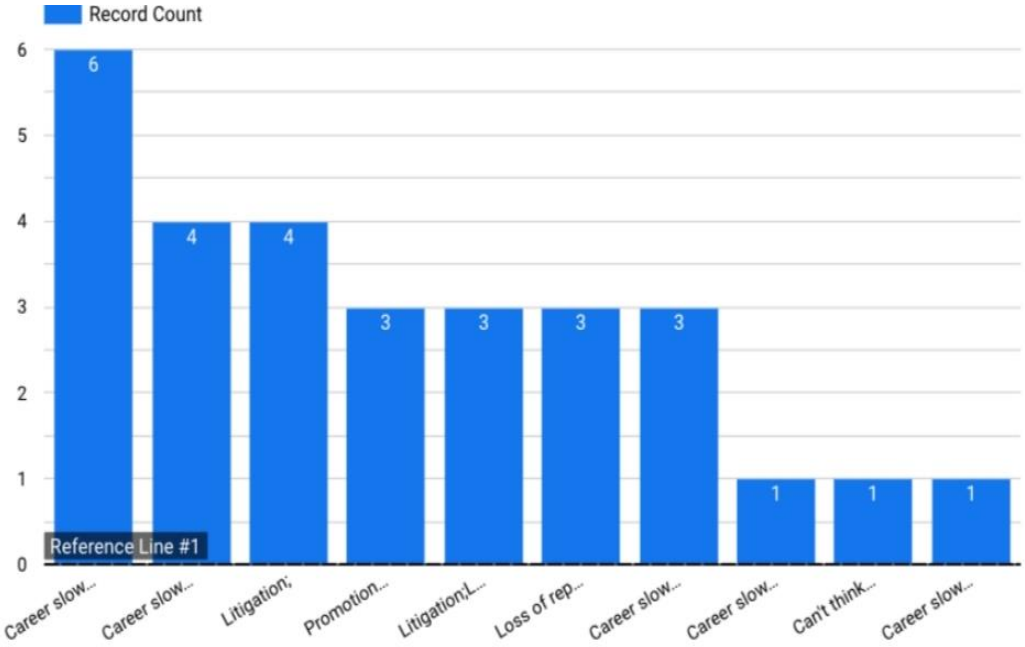

Figure 7: Risks Resulting From Stolen Data

4.1.9 Research Collaboration Experience

Table 2 answers the Research Question 3: "What has been the foreign research collaboration experience of scientists?". It shows that 65% of the respondents have had foreign research collaboration, with only 26% hosting foreign collaborators at their institution. A large percentage (79%) have shared their data and methods with collaborators, but only 19% have had their research laboratory replicated in a foreign institution. About 55% have travelled outside the country for research purposes

Table 2: Research Collaboration Experience

	Yes	No	Maybe
Have you been involved in foreign research collaboration?	31	16	
Have you hosted foreign collaborators/students as part of your research?	12	35	
Have you been involved in research programs that require you to share your data, methods and knowledge?	37	5	5
Have you been involved in research that attributes the awards, patents, publications to a foreign institution?	16	26	5
Do you have third-party funding that replicated your research lab work in a foreign institution?	9	38	
Have you travelled outside Nigeria to conduct research?	26	21	

4.1.10 Purpose of Research

The data in Table 3 confirms that 64% respondents are not aware of dual use research and 94% have not received any training on dual use concerns before. A large number (32) are interested in more information related to dual use. A minor percentage of respondents are convinced that their research can be used for dual use. About half of the respondents (53%) do not know how their research can be put to dual use. Hence, Table 3 answers the Research Question 1: "Do researchers know if their work has dual-use?" 53% of the researchers are not aware if their work has dual use while a larger percentage are not even conversant with dual use research.

Table 3: Research Purpose

	Yes	**No**	**Maybe**
Have you heard of the term "dual-use research" before?	9	30	8
Have you received any formal training or education on dual-use research issues?	3	44	
Do you believe there is a need for more educational resources on dual-use research for researchers in your field?	32	2	13
Can your research be used for both good and evil?	12	35	
Do you know how your research can be used for good or evil?	22	25	
Can your research be commercialized directly?	26	4	17

4.1.11 Results Showing Computing Skills

From Table 4, majority of the respondents do not use the institutions ICT infrastructure for research and have not been trained to respond to cyber attacks. A high percentage use antivirus as a data protection tool. About 70% of the respondents have not experienced ICT failure or cyber attacks.

Table 4: Computing Skills

	Yes	**No**	**Maybe**
Do you use the University ICT infrastructure for research?	17	30	
Have you ever lost research data due to ICT infrastructure failure or attack?	14	33	
Do you have an antivirus?	38	9	
Have you been trained on how to identify and avoid cyber attacks?	10	37	

4.1.12 Risk Identification

Table 5 shows that the respondents who engage in research and do not know all the terms of agreement are 60%. Many respondents (85%) do not check their research for dual use purpose while only 45% of respondents do a due diligence on potential collaborators or students.

Table 5: Risk Identification

	Yes	**No**
Do you have access to the research collaboration agreement or MOU signed for your research?	19	28
Do you assess the potential risks associated with your own research projects in terms of dual-use concerns?	7	40
Do you assess(investigate) potential research collaborators/students before engagement?	26	21

Results from this study have shown that the science-based researchers that were investigated have little or no knowledge about dual-use technology and also lack the skills to counter cyber attacks. Minimal due diligence of research collaborators and local storage of research data were also observed. Hence, there is the need to sensitize the research community about incorporating research security practices and how to handle dual use research in order to forestall the activities of malign actors.

4.2 Knowledge Modelling of the Survey Through Data Boosting Using Machine Learning Techniques

The performance of the dataset from both oversampling datasets were used to build XGBoost. The results of SMOTE oversampling show a better performance than Random oversampling in all the metrics, including accuracy, precision, recall and F1 Score as shown in Figure 8. The experiment was similarly repeated for traditional Random Forest (RF) to balance the investigation.

	XGBoost	Accuracy	Precision	Recall	F1 Score
0	SMOTE Oversampling	0.914400	0.933527	0.895125	0.901873
1	Random Oversampling	0.871625	0.781319	0.740157	0.727811

Figure 8: Results of XGBoost for the Two Oversampling Techniques

The performance of SMOTE oversampling informed the use of the synthetic dataset generated from the technique to build three models, as depicted in Figure 9. The ensemble technique slightly outperforms the Support Vector Machine (SVM) with an accuracy of classification of 0.914 and 0.9123 against 0.9107 for XGBoost, RF and SVM, respectively. This alludes to many researchers claiming that ensemble methods are better models when it comes to classification.

The precision result from XGBoost demonstrates 93.35% of predictions of total instances of those researchers who are aware of dual-use research, while the model failed to predict accurately 6.7%, giving a false call. However, SVM had a slight improvement ahead of XGBoost with 93.80%, while traditional RF recorded a low percentage of precision value with a correct prediction of 91%. It has a higher false positive of 9%, which may impact our assumption about researchers' understanding of dual-use research.

SMOTE Oversampling

	Accuracy	Precision	Recall	F1 Score
XGBoost (XGB) Classifier	0.9144	0.933527	0.895125	0.901873
Random Forest (RF)	0.9123	0.929774	0.897375	0.900616
Support Vector Machine (SVM)	0.9107	0.938069	0.888463	0.901009

Figure 9: SMOTE Oversampling of XGB, RF and SVM

Figures 10 and 11 show the importance of the features of XGBoost and Random Forest.

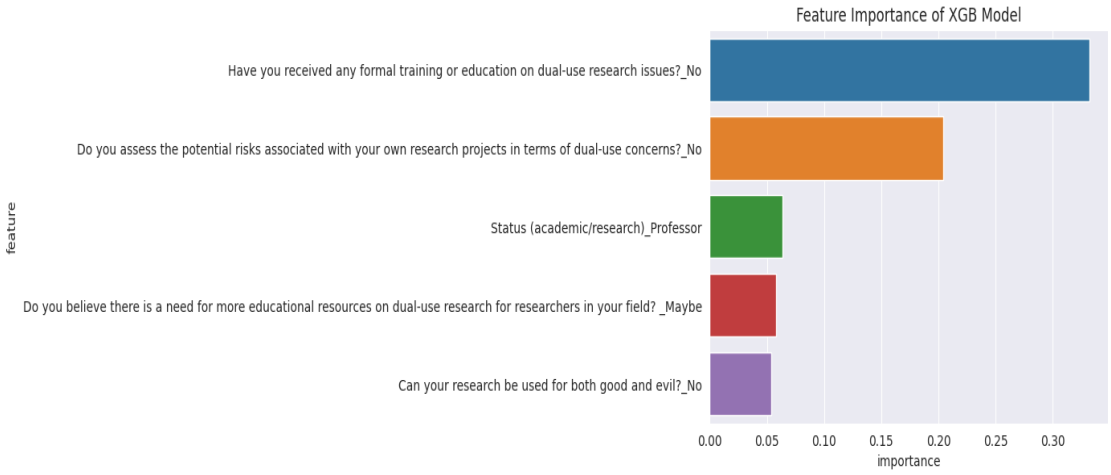

Figure 10: Feature Importance of XGBoost Model

The feature importance is significant to selecting the best features when building a model. It gives more insight into the dataset and provides information on the feature quality that contributes most to the final prediction. The less important features were separated and ignored during the model's training. The XGBoost model was better in all the metrics. Hence, its feature's importance is relevant to this study. The model prioritizes formal training of researchers on dual-use research as a major feature when performing classification of researchers in this work as seen in Figure 10. Understanding the potential risk associated with the researcher's work was also a main factor that a premium should be placed on.

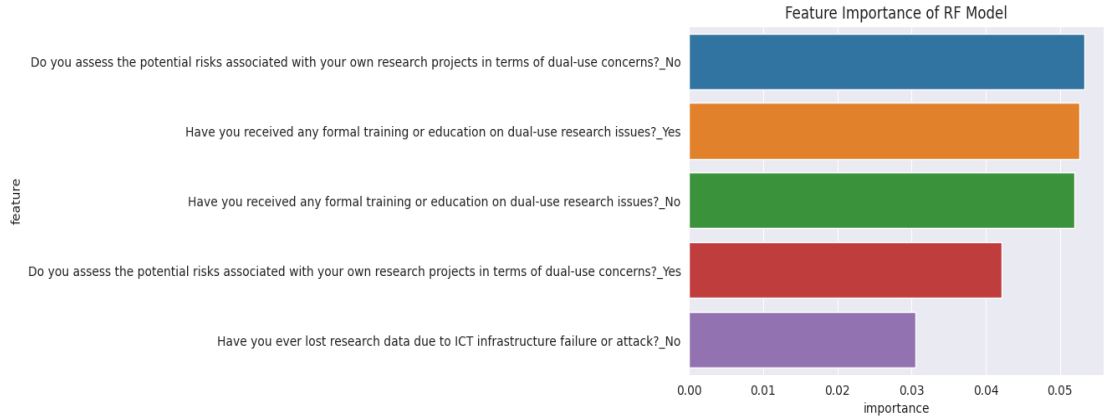

Figure 11: Feature Importance of Random Forest Model

Similarly, traditional Random Forest placed high regard on understanding the potential risk of research carried out by researchers, as illustrated in Figure 11. The formal training of researchers also contributes immensely to the model building.

5. CONCLUSION

In conclusion, the knowledge discovery initiative of this study serves as a crucial step toward better understanding the landscape of research security practices among scientists, providing actionable insights that can enhance data protection, security awareness initiative and foster a more secured environment for scientific exploration and collaboration. It will also guide the scientific community in strengthening its defenses against threats to research integrity.

6. ACKNOWLEDGMENTS

We appreciate Sandia National Laboratories (SNL) and The U.S. Department of State for training us on Research Security and for sponsoring this research.

7. REFERENCES

[1] U.S. Office of Science and Technology Policy, 2022

[2] Sandia National Laboratories (SNL). 2023. "Research Security Webinar Series: An Introduction. Powerpoint Slide, page 31.

[3] Government of Canada. (2023). Why Safeguard Your Research. (https://science.gc.ca/site/science/en/safeguarding-your-research/general-information-research-security/why-safeguard-your-research)

[4] Security and Integrity of the Global Research Ecosystem (SIGRE) Working Group. 2023. G7 Best Practices For Secure & Open Research.https://www8.cao.go.jp/cstp/kokusaiteki/g7_2023/2023_bestpracticepaper.pdf

[5] Priyad H. (2023) What is machine learning and how does it work? Online Tutorial retrieved from https://www.simplilearn.com/tutorials/machine-learning-tutorial/what-is-machine-learning

[6] Ayorinde, I. T. and Osofisan, A. O. (2010). Application of Artificial Neural Networks in Classifying Medical Database. Journal of Science Research Vol. 9: 12-18.

[7] Ayinla B. I. (2023). Index Mapped Ordinal Encoding Method for Federated Machine Learning in Crime Detection. University of Ibadan Journal of Science and Logics in ICT Research (UIJSLICTR) Vol. 9. No. 1:6-20.

[8] Wilson P. F. (2020). Academic Fraud: Solving the Crisis in Modern Academia. Exchanges: The Interdisciplinary Research Journal· June 2020. 7(3), pp. 14-44 DOI:10.31273/eirj.v7i3.546

[9] National Co-ordinating Centre for Public Engagement (NCCPE). (2020). www.publicengagement.ac.uk

[10] Linkedin (2023). How can you collaborate with other researchers while minimizing risks? https://www.linkedin.com/advice/3/how-can-you-collaborate-other-researchers-while-minimizing

[11] Texas Tech University (TTU). (2023). What are Some Best Practices for Researchers to Protect Research Data? https://www.ttu.edu/it4faculty/research/

[12] Ayorinde, I. T. and Badmos, Z. O. (2019). Development of Deep Learning Model on Mushroom Dataset towards Classifying Poisonous Mushroom with Feature Selection. In Odumuyiwa, V., Onifade, O., David, A. and Uwadia, C. (Eds.). Transition from Observation to Knowledge to Intelligence. Lagos, ISKO-Nigeria. 225-237pp. ISBN: 978-978-976-000-8.