Research Methodology Final Exam – Group Work

1. Submit the final paper of your group research work using the IEEE template / the template from the journal/conference of your choice as agreed with your lecturer
2. The paper in DOCX format named “Final Paper – K15.docx” can be found in the zip folder
3. The paper in PDF format named “Final Paper – K15.pdf” can be found in the zip folder
4. Each member contributions/job description:

* Bryan Orville Audric

Choosing the appropriate model

Working on dataset collection and Preprocessing

Choosing the appropriate model

Writing the first version of Introduction, Related Work, Materials & Method, and Result & Discussion

Writing and Revising References

Revising Result & Discussion, Materials & Method, Conclusion and Abstract

Doing a final check of the paper before submitting

* Edrico Putra Pramana

Working on Feature Selection, Modelling and Evaluation of the model

Adding Some necessary Reference

Writing the first version of Conclusion and Abstract

Completing Materials and Method

1. Include proof of submission of your paper to the international conference / journal along with the approval of your lecturer. Mention all the submission information detail:
   1. Submission Details:

* The conference/ journal website link:

2024 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT)

<https://comnetsat.org/>

* Organizer (for international conference) or Publisher (for international journal):

Communication, Networks and Satellite (COMNETSAT)

* Scope:

Big Data, Data Representation and Visualization

* Date or Submission date (International Conference):



Venue Location (International Conference) : Lombok – Indonesia

Further details can be found in <https://comnetsat.org/submissions/>

* Keynote speaker (for international conference):
  + Prof. Mohamed-Slim Alouini

Computer, Electrical, and Mathematical Science and Engineering (CEMSE) Division King Abdullah University of Science and Technology (KAUST) Thuwal, Makkah Province, Saudi Arabia.

* + Prof. Abbas Jamalipour

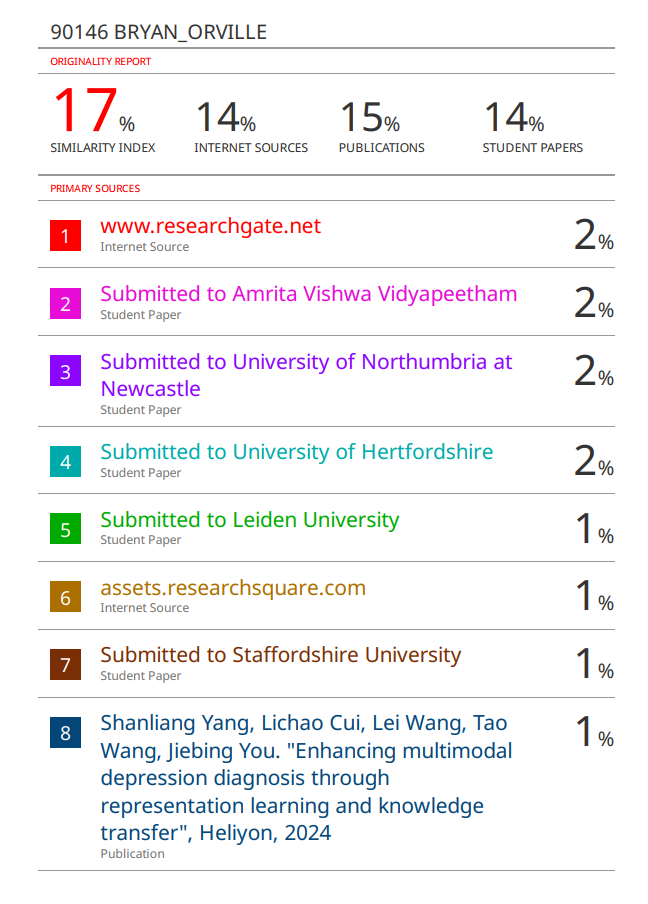
Professor of Ubiquitous Mobile Networking School of Electrical and Information Engineering, The University of Sydney, Australia

* + Prof. Nirwan Ansari

Professor of Electrical and Computer Engineering at the New Jersey Institute of Technology (NJIT)

Further details can be found in <https://comnetsat.org/speakers/>

* Turnitin similarity result (must be lower than 20% or as agreed with your lecturer). You can ask your own lecturer for this / Request to LKC at <https://bit.ly/libraryturnitin> / etc:

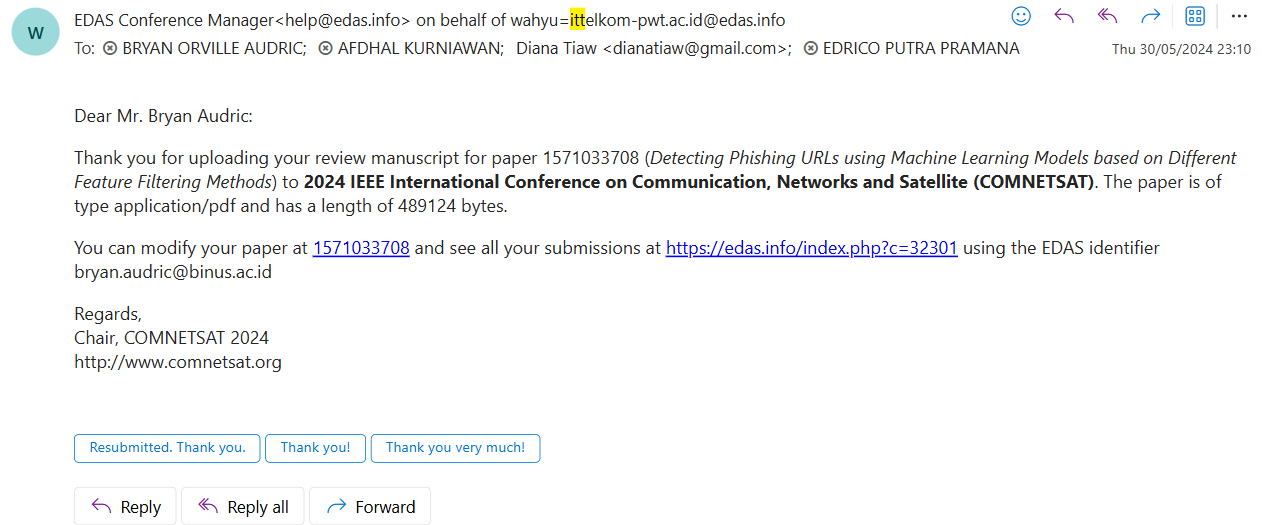


Further details can be found in the file named “Turnitin – K15.pdf” located inside the zip folder

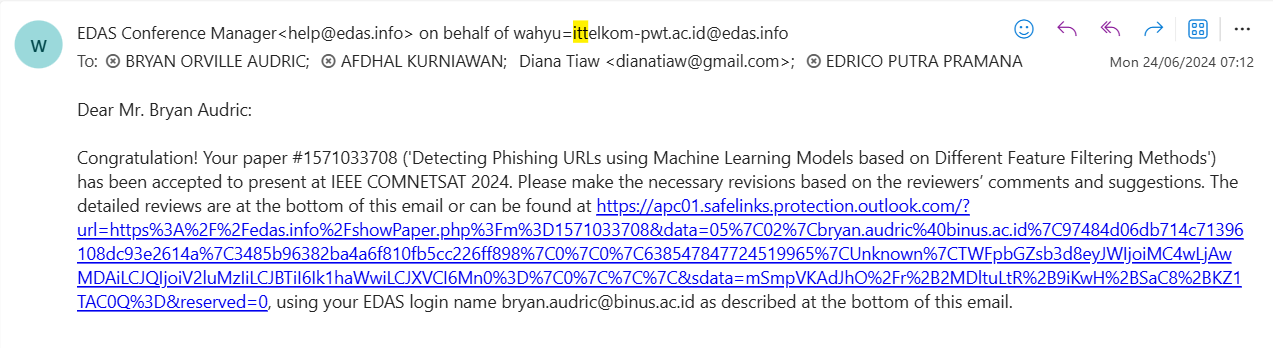
Note: The format of the paper submitted to Turnitin was in ICCSCI format while the format submitted to Conference is in IEEE format. The content of the paper is the same, it is just in different format.

* 1. Paper Submission Proof

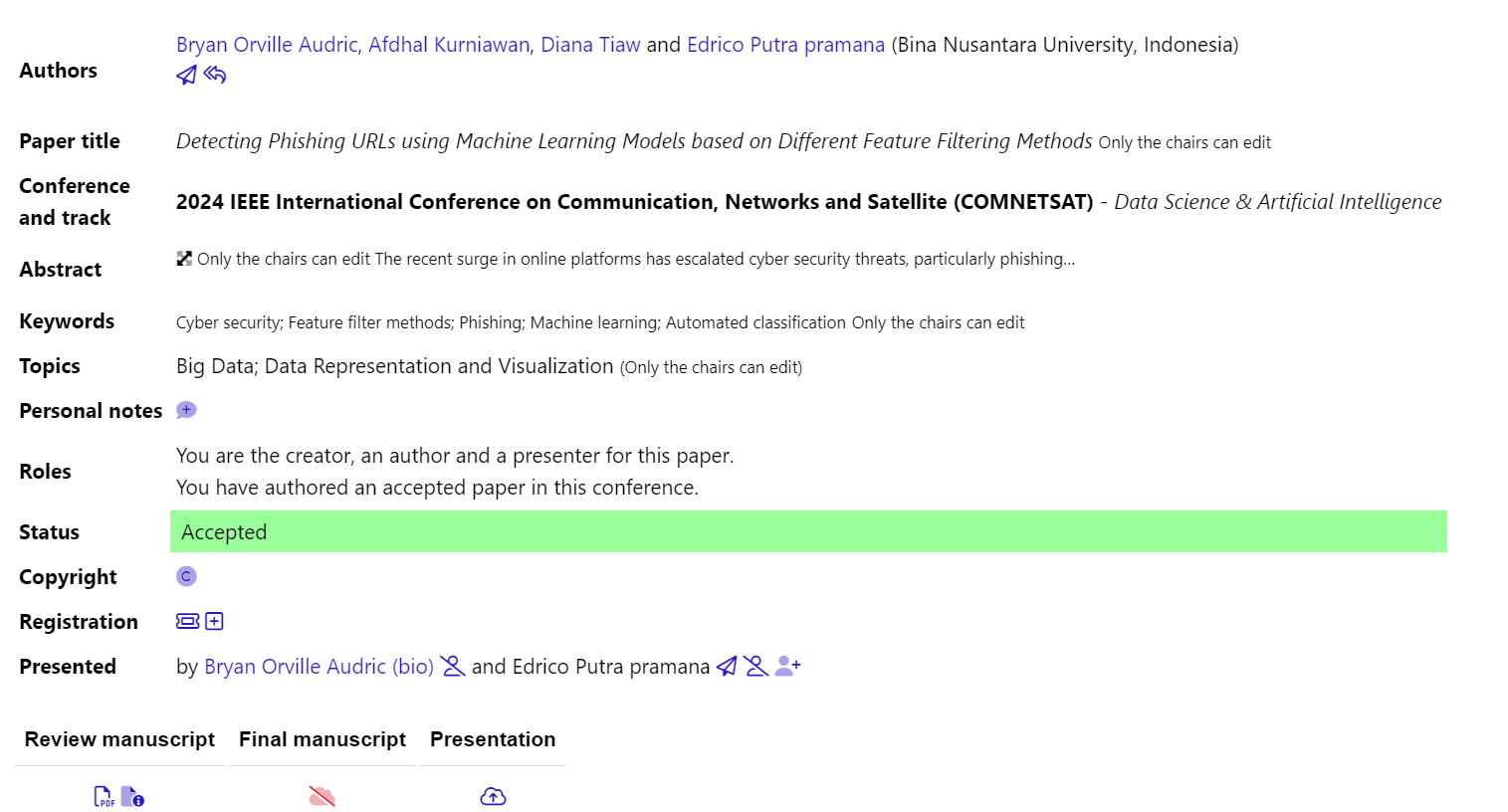
Submission Proof (Screenshot of the email)



Paper Acceptance Proof (Screenshot of the email)

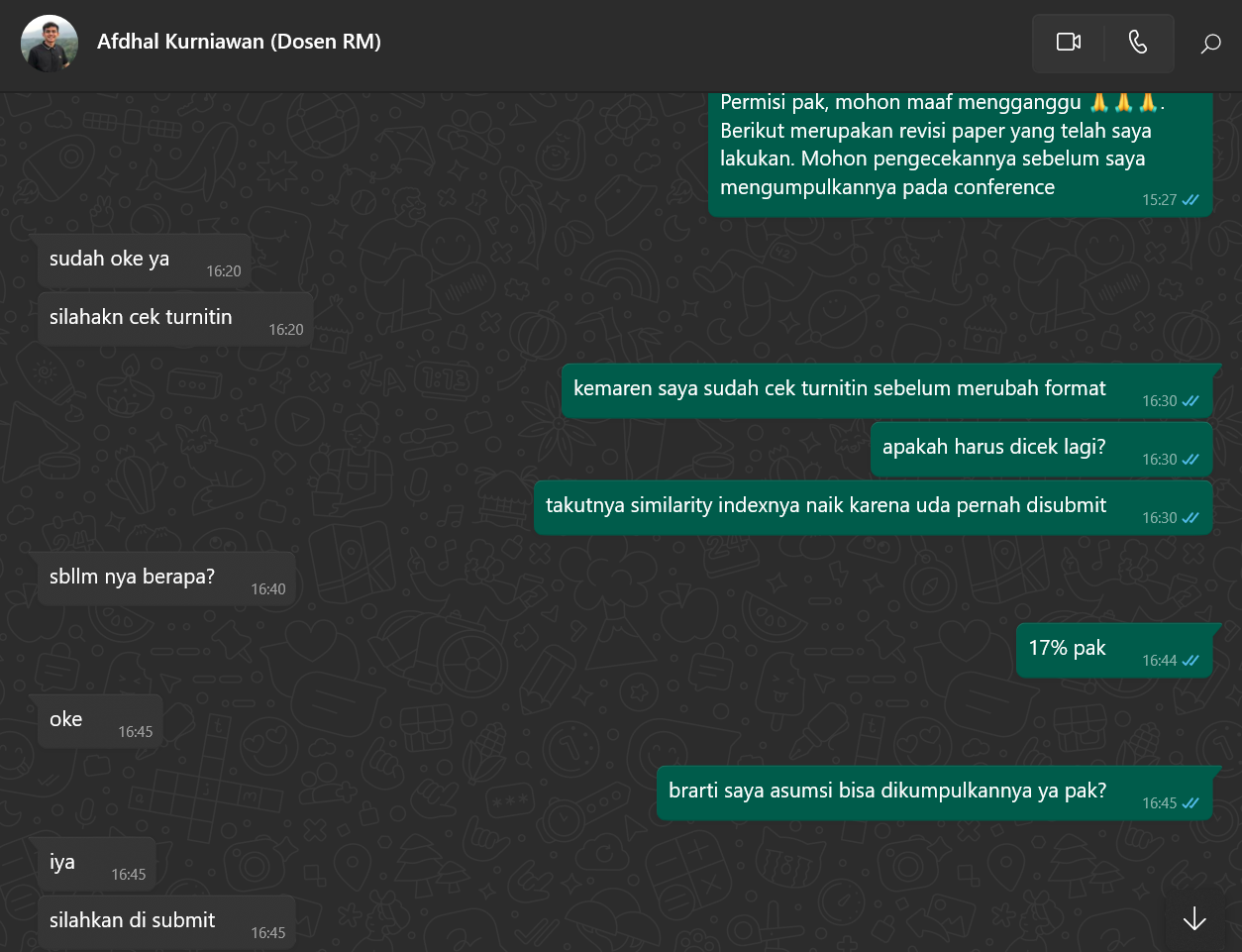


Screenshot of the EDAS website



* 1. Lecturer approval proof of your paper (screenshot of the approval email / chat / other forms as agreed with your lecturer):

Lecturer Approval Proof (Screenshot of chat)



* 1. Turnitin Similarity Report (Turnitin – GroupName.pdf):

The Turnitin Similarity report named “Turnitin – K15.pdf” can be found in the zip folder

1. As a part of scientific communication, your group must create a research trailer. The duration of the video is 1-3 minutes (no less and no more), consisting of your background, your methodology, your result, and your conclusion. As the target audience is for the public (think of it as communicating through social media), you may edit your video creatively to make it more interesting.​

Research trailer video link: <https://www.youtube.com/watch?v=pGW8sVUrNlA>

1. To simulate the presentation during the conference, together with your group members, youmust create a presentation video (landscape) of your publication (act as if you’re presenting in an international conference, use formal attire). The video duration is 10 minutes Max. In your first slide of the powerpoint, make sure to include your title, authors’ name and affiliation, BINUS Computer Science Logo, and the conference/journal of choice. You are still required to make the video even if you submit to a journal!
   1. The script of your presentation

Edrico Putra Pramana’s Script:

Slide – 1

Hello everyone, my name is Edrico Putra Pramana, and this is my partner Bryan Orville Audric. Today we’ll be presenting to you our research on Detecting Phishing URLs Using Machine Learning Models Based on Different Feature Filtering Methods. We will be presenting this on International Conference on Communication, Networks, and Satellite. Before we start presenting, I want to thank you my lecturers in BINUS University, Mrs. Diana and Mr. Afdhal Kurniawan for the guidance during the research process.

Slide – 2

Skip

Slide – 3

The proliferatioin of online platforms has heightened cyber threats, particularly phishing and malware attacks aimed at stealing sensitive data. To address this, researchers have increasingly use machine learning models to automate phishing detection.

Slide – 4

Our objective in this research is to conduct an evaluation and comparison among several feature filters used in phishing detection using machine learning models.

Slide – 5

There quite a few related works that we used during this research, the first one being Deteksi Website Phishing Menggunakan Teknik Filter Pada Model Machine Learning this one uses Pearson Correlation filter feature method and Random forest as the model, the second one is Towards Benchmark Datasets for Machine Learning Based Website Phishing Detection An Experimental study, this one uses Chi Square filter feature method, and the third one is Phishing Detection Using Machine Learning Techniques, this one uses Gradient Boosting Algorithm specifically XGBoosting. These three achieved quite high accuracy in terms of detecting phishing URLs.

Slide – 6

This is the overview of our research.

Slide – 7

The dataset is taken from kaggle, it contains 651.000 URLs. There are 4 types of URLs, benign, Defacement, Phishing, and Malware, benign URL is a safe URL, defacement URL is a URL that mimics original URL with the intent to trick the users, Phishing URL is a URL for a phishing attempt, and malware URL is URL with malware in it.

Slide – 8 - 10

This is the feature list; we have 25 features that we use during this research.

Bryan Orville Audric’s Script:

Slide – 11

So, I will be continuing the presentation. This is feature-selection’s part. For this process, we are using Pearson correlation to help filter out features. Pearson Correlation is a statistical metric that basically calculates the correlation coefficient of every features. The feature with highest correlation coefficient will be taken as the subset of features.

Slide – 12

Similar to Pearson correlation, chi square is a statistical metric that basically calculate how relevant a feature is to the target value.

Slide – 13

Linear Discrimant Analysis is a naturally a dimensionality reduction technique, but we are using this method to help calculate the eigenvalues of every feature and then we grab the feature with highest eigenvalues

Slide – 14

So, this is the first machine learning models, we are using random forest. It is basically an ensemble machine learning model. It begins with bootstrapping by producing different tree with different subset of features and ends with majority voting where the majority answer of the tree will be the prediction of this random forest model

Slide – 15

This is XGBoost. It is a gradient boosting algorithm that will be used to predict phishing. It basically just building a tree sequentially and then it will correct the error of the previous tree. At the end of it, we will get a tree that can predict accurately

Slide – 16

This is LightGBM. It is just an improvement of XGBoost and it is more efficient by growing its tree leaf wise compared to XGBoost that grows its tree level wise

Slide – 17

This is the evaluation metrics used for this research. The first one is accuracy, the second one is f1 score, the third one is precision, and the fourth one is recall. The equation is shown in the slide

Slide – 18

So, this is the feature that we got from Pearson correlation. As you can see, we got 8 features, and all features have an absolute value of correlation coefficient above 0.1

Slide = 19

This one is chi square’s result. We calculate every feature’s chi square value and then we picked 10 features with the highest chi square value

Slide – 20

This one is Linear Discriminant Analysis’s feature. It was able to produce 4 subsets of features since there were 4 classes. Amongst those 4 subsets of features, the third subset of features achieve the highest accuracy. The third subset of features is shown in the slides

Slide – 21

This is the Pearson correlation’s result. It averages in 94% accuracy across all models, with Random Forest achieving the highest accuracy of 95%

Slide – 22

This is the Chi Square’s result. It averages in 91% accuracy across all models, with Random Forest achieving the highest accuracy of 92%

Slide – 23

This is the Linear Discriminant Analysis’s result. It averages in 93% accuracy across all models, with Random Forest achieving the highest accuracy of 94%

Slide – 24

As mentioned before, Pearson correlation’s result has an average accuracy of 94%, in which Random Forest has the highest accuracy, similar to other filtration method

Slide – 25

The conclusion of this research is that we can indicate that Pearson correlation in this particular dataset perform better compared to other filtration methods, achieving the highest accuracy of 95% with the help of machine learning models

Slide – 26

I would like to suggest that there is a limitation of this study. We feel that the lack of phishing and malware URL causes low precision, recall and f1 score as you can see over here. It is lower compared to other categories. This is why that the accuracy (slide - 21) tops at 94%

Slide – 27

I think that’s all for our presentation. Thank you very much

* 1. The Link to the slide (PPT/PDF) you used in editable format.

Slide OneDrive link in PPT format: [Project - Presentation Slides - K15.pptx](https://binusianorg-my.sharepoint.com/personal/bryan_audric_binus_ac_id/_layouts/15/guestaccess.aspx?share=EZfdG1z-U3RMq5gT8fiDVkYBefDqn8702z4c5oRMZFqB-g&e=dIlHfV)

* 1. The Link to your video (Upload it either in YouTube or your Personal Drive).

Presentation video link: <https://youtu.be/hmrgJcosH6c>