# 

# Introduction

* **Khanh viết định nghĩa vấn đề *(Bảo hiểm là gì, bảo hiểm xe?)***

Insurance is a contract where an insurer agrees, for a premium, to provide monetary

coverage to the insured party in case of a defined loss resulting from uncertain events, after a formal claim is filed by the claimant party (Viaene, S., & Dedene, G., 2004).

Viaene, S., & Dedene, G. (2004). Insurance fraud: issues and challenges. The Geneva Papers on Risk and Insurance - Issues and Practice, 29(2), 313–333. https://doi.org/10.1111/j.1468-0440.2004.00290.x

Đền bù hợp đồng bảo hiểm

Tỉ lệ gian lận bảo hiểm

Xác định như thế nào:

* Truyền thống
* Ảnh hưởng doanh thu như thế nào
* Ảnh hưởng chính đến kế hoạch kinh doanh

Phương pháp trước giờ, ngắn gọn

* **Lộc** điền giúp chị các mô hình dưới đây và kết quả em mong muốn của nó để chị tìm ref phù hợp viết nha

+ Các hướng tiếp cận truyền thống, ưu và nhược điểm

…

+ Các hướng tiếp cận hiện đại, ưu và nhược điểm

…

* Tại sao mình làm, làm gì, đóng góp (**Chi với Ngọc** tự chọn chia phần làm). Mỗi đoạn dưới đây viết từ 10 đến 20 dòng là được. Phần dưới này thì ko cần ref. Mỗi đoạn điền bên dưới lun, đừng xóa cái yêu cầu của chị.

## Dataset:

(chưa mô tả cụ thể nha, ghi bộ của mình nó update so với mấy bài trước như thế nào, phù hợp với thị trường, xu hướng hiện tại như thế nào) (Chi)

In this dataset, we utilized models to predict insurance claims for automobile insurance. The dataset comprises 2015 insurance claims, which were split into datasets for training, hyperparameter tuning, and testing model accuracy. This serves as a foundation to develop models for predicting automobile insurance claims, evaluate their accuracy, and understand the influencing factors behind these claims.

An important aspect of this dataset is the presence of missing values and noise, necessitating preprocessing before analyzing and encoding these variables to enhance the performance of machine learning models. Furthermore, the dataset reveals a high correlation among variables such as vehicle value, claim amount, accident count, and some variables with skewed distributions, like claim amount and accident count. This requires addressing multicollinearity and removing redundant variables to make distributions closer to normal.

The diversity of the data makes it valuable for developing predictive models and fraud classification. This can help insurance companies gain a better understanding of the determinants of insurance claims and enhance fraud detection capabilities.

## Method (mình phân tích những model gì, kết quả ntn) (Chi)

During the training process, we built both unsupervised and supervised machine learning models. For the unsupervised model, the team attempted to construct a deep learning Neural Network. For the supervised models, the team selected the following models such as KNeighborsClassifier, GaussianNB, SVC, MLPClassifier, GradientBoostingClassifier, LogisticRegression, LinearDiscriminantAnalysis, DecisionTreeClassifier, and RandomForestClassifier. The results indicated that the Random Forest model performed the best.

## Contribution (kết quả này đóng góp gì được trong thực tế) (Ngọc)

In our view, the results of this study make a significant contribution to the field of automobile insurance. The newly developed model in the research demonstrates a more effective capability in predicting auto insurance fraud compared to previous models. This aids auto insurance companies in more accurately understanding and managing financial risks, thus avoiding losses due to fraudulent activities. Consequently, insurance companies can formulate more precise strategies based on the market's actual conditions.By utilizing a high-performance model as presented in this study, insurance businesses can proactively prevent fraudulent cases and minimize financial losses. The model is designed to be flexible and applicable in various market scenarios, showcasing its generalizability. Therefore, we assess that the applicability of this model extends beyond the auto insurance market and could be implemented in related industries such as financial fraud in banking, healthcare insurance, life insurance, and more.The outcomes of the model not only bring cost savings and efficiency benefits but also assist insurance enterprises in optimizing long-term decisions and strategies. This contributes to mitigating negative aspects within the automobile insurance industry and potentially other related fields.

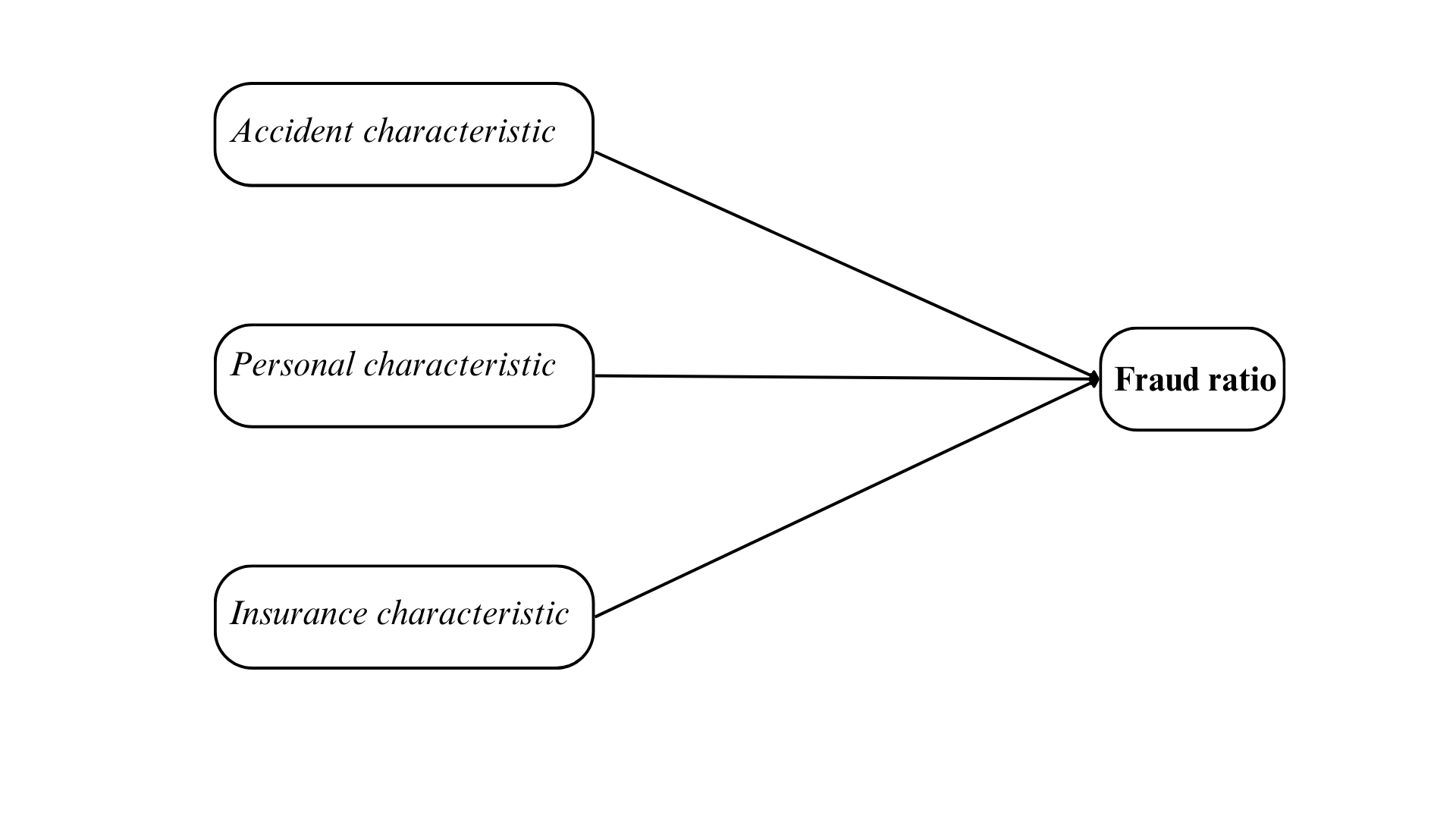
## Về mô tả sơ cấu trúc bài (ngắn gọn hoi) (Ngọc)

The research structure has been designed in a logical and flexible manner to enhance readability for readers. Part one introduces and frames the research problem and objectives. Here, we present the traditional and modern approaches to achieving the research goals. We then provide an analysis of the advantages and disadvantages of these approaches and introduce the new methods and models developed by our team. Part two provides a detailed description of the dataset, limitations in the data, previous studies on fraud issues in insurance, and outlines the testing and implementation methods used in our model. Part three presents the results of the model. Part four discusses the findings and the significance of the research. Finally, in the concluding section, we summarize the research and propose directions for future research.

# Materials and Methods

## Về dataset

* **Lộc** đưa vào bảng thống kê mô tả các biến



|  |  |
| --- | --- |
| **Nhóm** | **Biến quan sát** |
| *Accident characteristic* | insured\_zip  capital-gains  capital-loss  incident\_date  incident\_type  collision\_type  incident\_severity  authorities\_contacted  incident\_state  incident\_city  incident\_location  incident\_hour\_of\_the\_day  number\_of\_vehicles\_involved  property\_damage  bodily\_injuries  witnesses  police\_report\_available |
| *Personal characteristic* | age  insured\_sex  insured\_education\_level  insured\_occupation  insured\_hobbies  insured\_relationship |
| *Insurance characteristic* | months\_as\_customer  policy\_number  policy\_bind\_date  policy\_state  policy\_csl  policy\_deductable  policy\_annual\_premium  umbrella\_limit  capital-gains  capital-loss  total\_claim\_amount  injury\_claim  property\_claim  vehicle\_claim  auto\_make  auto\_model  auto\_year  fraud\_reported |

* **Chi** viết mô tả dataset

The input of the dataset consists of 41 columns and 1000 lines. The information of the dataset includes:

* months\_as\_customer: The number of months the applicant has been a customer with the insurance company.
* Age: The age of the insured person.
* policy\_number: Insurance ID.
* policy\_bind\_date: Start date of the insurance contract.
* policy\_state: The state where the insurance policy is issued.
* policy\_csl: The Single Combination Limit (CSL) indicates the maximum amount the policy will pay per incident. Example of the “250/500” CSL package in the article: The maximum limit for human damage is $250,000.
* The maximum limit for property damage is $500,000.
* policy\_deductible: The amount the policyholder has to pay himself before the insurance company starts paying. For example, a customer has a car insurance policy with a deductible of $500. If a customer has an accident and the cost of repairs is $2,000, they'll have to pay the first $500 from their pockets, and then the insurance will pay the rest of the $1,500.
* policy\_annual\_premium: The amount of the annual fee for the insurance policy.
* umrella\_limit: An additional type of insurance is designed to provide a higher coverage limit after the basic policy limit has reached its maximum. For example, if a customer has a $250,000 car liability policy and an umbrella policy with a $1 million limit, when an accident occurs and the cost of compensation is $800,000, the car policy will pay the first $25,000, then the umbrellas policy pays the rest of the $550,000.
* Insured\_zip: The ZIP code of the insured.
* Insured\_sex: The sex of the insured.
* Insured\_education\_level: The educational qualifications of the insured.
* Insured\_occupation: The occupation of the insured.
* Insured\_hobbies: The preferences of the insured.
* Insured\_relationship: The relationship of the insured person with the policyholder.
* capital-gains: The capital gains the insured receives against the amount of the annual premium for the insurance policy.
* capital-loss: The capital loss of the insured against the amount of the annual fee for the insurance policy.
* incident\_date: The date the incident occurred.
* incident\_type: Type of incident (e.g. collision, theft).
* collision\_type: Type of collision if any.
* incident\_severity: The severity of the incident.
* authorities\_contacted: Which authorities were contacted when the incident occurred.
* incident\_state: State where the incident occurred.
* incident\_city: The city where the incident occurred.
* incident\_location: The specific location of the incident.
* incident\_hour\_of\_the\_day: The time of the day when the incident occurred.
* number\_of\_vehicles\_involved: The number of vehicles involved in the incident.
* property\_damage: Indicates if there is any property damage.
* bodily\_injuries: Number of injuries caused by incidents.
* Witnesses: The number of witnesses for the incident.
* police\_report\_available: Indicates whether there is a police report.
* total\_claim\_amount: The total amount subscribed for the incident.
* Injury\_claim: Registration for personal injury.
* property\_claim: The amount registered for property damage.
* vehicle\_claim: The amount registered for car damage.
* Auto\_make: The manufacturer of the vehicle is insured.
* Auto\_model: The model of the vehicle is covered.
* Auto\_year: Year of manufacture of the vehicle is covered.
* fraud\_reported: Indicates whether the application is to be fraudulent.
* \_c39: This column appears to be incomplete or a spacecraft with no clear description.

Contains information about the data that the team uses to analyze or build a model. Here, the input of the data set is specifically about car insurance claims such as:

* Information about the owner: age, gender, educational status, occupation, interests, relationships, etc.
* Information about the car: vehicle type, value, car manufacturer, year of manufacture, amount registered for car damage...
* Information about the accident: time, location, type, type of collision, severity, number of witnesses, etc.

The output of the data set was intended to detect insurance fraud, which has caused no less damage to insurance companies and society as a whole. Detection of insurance fraud is a branch of detection of financial fraud. Specifically at the output here is to ***Require the compensation contracts to be valid or fraudulent***.

The problem to solve that the team proposes is to classify the complaint contract as accurate or fraudulent. In this context, the detection of insurance fraud not only reduces unnecessary costs, but also ensures fairness and transparency in the process of compensation, protecting the rights of both the insurer and the company. If it lasts, it will affect the valuation strategy and the socio-economic benefits of the insurance company. So, on the basis of this data, state regulators make appropriate policies.

* **Ngọc viết limit của dataset (kiểu có hạn chế gì trong bảng dữ liệu này)**

In our research on the dataset of Automobile Insurance Claims, we have identified several limitations. Firstly, we attempted to update and expand the dataset to better reflect the current market reality. However, there are still discrepancies in the data compared to the actual market, particularly as the dataset was obtained from the year 2015 in 7 states of the United States. This limitation may impact the comprehensiveness and overall representation of the auto insurance industry. Secondly, the absence of information or important variables in the dataset has the potential to affect the model's predictive capabilities. We addressed this issue through data exploration, analysis, and employing various data processing strategies.Thirdly, the dataset's large size and the limited compatibility and correlation among variables may pose challenges. This impacts data quality and could lead to overfitting if not managed carefully. Fourthly, there is an imbalance in the data between predicting fraud and non-fraud cases. This imbalance reduces the predictive performance of the model. These limitations should be carefully considered when evaluating and applying the model results.

**Về nhận xét EDA**

* **Chi** dịch bên bài cũ qua được á. Phần tiền xử lý dữ liệu để trong part này lun. **(6:32 - 22/1)**

After conducting data analysis and exploration, in order for the topic to be conducted to the best possible performance, the team drew some important conclusions and transformed the data as follows:

The "c\_39" column contains all the empty values and does not mean anything to the data set, which is why the team removed it from the dataset. Moreover, the data set does not show empty values, but when we examine more closely, we can see the question marks, which do not mean any of the values in the dataset, and we can find that it is the defective values of the data suite that need to be dealt with immediately.

When it comes to analyzing the classification variables in the data set, the team notices classification values but there are many unique values in the dataset such as: the unnecessary client code for predicting that the team intends to remove that column, not just because the model does not receive the values of the "datetime" data type, which is why we extract the date, month, year from the two columns "incident\_date' and policy\_bind\_date" and proceed with removing the column.

Next, from the results obtained through experiments, the team found that classifying variables such as age, number of witnesses,... into groups would improve machine learning. So the team decided to split these attributes into groups and convert them into classification variables. At the same time, in some columns, the team has begun performing group separation to optimize input data for modeling.

On the part of commenting on the standard distribution and the exceptional values in the data set, the observer found that there were some exceptions, although they were insignificant. At the same time, some attributes have a non-standard distribution. The team proposed two strategies for data preprocessing. It's the processing of exceptional values using the IQR method, then using the MinMaxScaler to standardize the data. The second strategy focuses on safeguarding all data and implementing standardization with StandardScaler to serve the objectives of the topic.

When conducting a correlation factor analysis, the team observed values with correlations that fluctuated in the range from non-correlation to weak correlation. This suggests that appropriate data transformation measures need to be implemented to ensure that the data is optimized for the purpose of the topic.

After completing the data exploration process, we noticed a remarkable point that needed to be addressed. Out of 1,000 complaint claims, the percentage of contracts as fraudulent is very low, only close to 29% in the data set. This suggests that the data set is experiencing imbalance, a common condition when working with machine learning. That's why we need a method to deal with data imbalances accurately.

## Về method

* **Ngọc** **và Khanh** mỗi người viết 1 related work (ref) (Họ làm gì, dataset ntn, method j, kết quả ra sao, hạn chế cái gì)
* Lĩnh vực phát hiện

\* Cái này **Lộc** chọn dùm **ref hoặc mô hình** còn hạn chế để mng viết nha

Đối với bài báo về Insurance Fraud Detection : “Evidence from Artificial Intelligence and Machine Learning.” - tác giả Gomes et al.(2021).

The research article on Insurance Fraud Detection, titled "Evidence from Artificial Intelligence and Machine Learning" by Gomes et al. (2021), introduces a novel model for classifying fraudulent insurance contracts to prevent negative situations in the auto insurance industry. The model is developed based on a dataset collected in 2020 from the analytics department of a U.S. company (EXL Service). The dataset comprises 33 variables, including binary variables indicating the presence of fraud in insurance claims. The author reviews traditional research, particularly referencing studies using machine learning and data mining in fraud detection.

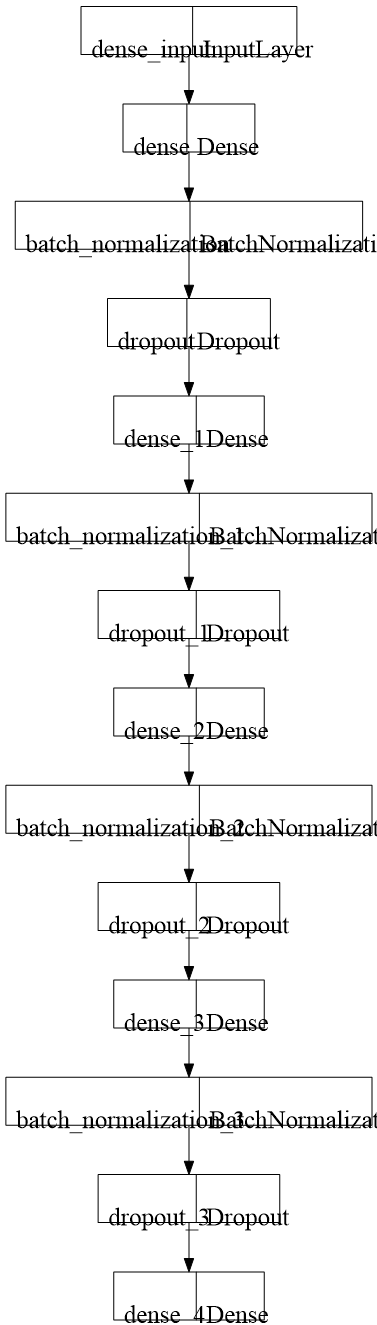
The research builds a machine learning model using three methods: Logistic Regression, Support Vector Machine, and Naive Bayes. Instead of relying on traditional models to identify important variables, the author uses the Boruta algorithm to determine the most critical variables for fraud detection. As a result, the author demonstrates that this model achieves high performance and practical applicability in the motorcycle insurance industry. SVM achieves the highest accuracy and specificity scores of 94% and 99.77%, respectively, while Naive Bayes attains the highest precision of 23.07%. However, concerning recall, F1-measure, and sensitivity, the Logistic Regression model outperforms with scores of 42.25%, 29.26%, and 42.25%, respectively. Overall, all models achieve high accuracy.

Compared with previous studies (Kumar et al., 2021), it is agreed that the Support Vector Machine is a promising predictive model. However, a considerable body of research suggests that Decision Tree, K-Nearest Neighbors, and Random Forest have superior performance over other machine learning and artificial intelligence models.

Nevertheless, the study has limitations as it focuses solely on auto insurance data in the United States. The dataset collected in 2020 may differ significantly from the current market reality. The research uses only three prediction models (LR, SVM, NB), and the performance of fraud detection might be higher when using a different model instead of the three mentioned. Additionally, the study is concentrated from the perspective of an auto insurance company, and it does not expand the viewpoint to policyholders or auto repair workshops, potentially reducing the comprehensiveness of the research.

* Xong r thì t sẽ viết 1 đoạn hạn chế chung của 2 cái ref đó, để dẫn xuống phần model dưới của mình.
* **Khanh** viết về các model, sẽ nhờ **Lộc** dán hình model sau

**Lộc** viết công thức mô hình dùm nhé!



# Results

# Discussion

# Conclusion & Future Works

support tham khảo: RD tốt nhất: [tại đây](https://www.mdpi.com/2227-9091/11/9/164)

chống tham khảo SVM tốt nhất: t[ại đây](https://pdf.sciencedirectassets.com/273213/1-s2.0-S0275531922X00032/1-s2.0-S0275531922001325/am.pdf?X-Amz-Security-Token=IQoJb3JpZ2luX2VjEKH%2F%2F%2F%2F%2F%2F%2F%2F%2F%2FwEaCXVzLWVhc3QtMSJGMEQCIEtNUdGlLB21ovbLE%2BHsDQkRg7sFCSuzjR2sY2Q38RU3AiA4DD9gEGFT3cJG98dVRuw04ZLTMWv27Gn%2FZfi3B06j8SqzBQhaEAUaDDA1OTAwMzU0Njg2NSIM7pr%2FVro5d%2BFZQChGKpAFJY3DvWDylDXRgRuR2jzXQ5vn4lOXv2zyn6w0LY0xMesf3Zxoo88HPOeRHvW26ZFqn%2BToOOAq3TE1SUscTSTpehax4FanOcIwop1GhWHVfkE1%2FH22kFlJvhkU7Xoyy0gKwLJDR2RPksgGYGXe4cpxUmV9HX%2FaC96VYJbkOP6zVRNz58xJBHPYQjDn%2BS2%2FGClC6iexYGVY%2Fyo7wqhp2NjMsQwA4x8LIYCiwCGpPVurBuZfzN2tykuEEcNu7sFDJnobl4lXn9w3PxgLXRQjcWG4bMGyKOKoYX%2BzdC2fIISuM6%2BvXgrWqY7AGfnL2K7R8SBxUq8P%2Fjr5lc04digMtfnAtX6tdj0g2iwmdGYhy%2BpUS%2FPhKw%2BGt3ygKI0utvvngOlTVk%2Fq3EhPJ5SlUTeCR%2Fw%2FqoapUGrzufdIEu5cTp04DaYylxOPTDBuL8gpashv2gZcihpAXnuYzEIzpKI4XOodgAKUd98%2F%2BIkaV91nfzmLL1gMqiVmMvsc%2Fyp12DJWX0eB8htUVxHxYlHKx3rxnpAh6lk%2FCO%2B3iUfrAuvqjtjqPTswCdwxI97ERB7tPr%2BxSkLLiiVHJ%2Fi92AcMbvxYAlFQKTy7yoBITNKQfjw1CtJAtshI1N8wVrjbAB63R%2BrT%2B32IKYAOLnrpFXinGQRJ23rMGMGCEIHEEEmafPTUPzEjI8FkMdCtN8oprUUSgkQ5sCKRTnihRUVPm%2BgCkXDB4ZWSX674tomjUG5bBkh7HM%2FMWQvpOlE3Mx%2FFko9%2Ftv%2BLmJNj6RPoxbudRo6Mwm0tGc1Hl%2FMVwfo8vsoavnhw%2BjwPc88N6Yp0LwM8ly4FJtmT8Aqk%2BFhYHO4rSDvMuCH8kVJjWmPEF1JAINnqeaCR1nq33AIwqu69rQY6sgGGgoqINPQFxtth%2FM6irB%2F82AiCtxbvYQ1bf%2BI%2BFzalWPijpfP7h%2Fx867%2F41a0qasOYeXe4Er%2F%2BiZkQHmtuWaIlIU%2FEWH2NgTfd1s1iVhHpoE32z4xrweX6XSqQLpI3QsZSn7NJMuH%2BQ%2FI8PUhSn728qhDx8a3xRm5UyrUex59n3cDzyfugFOZfmbQFkPg4NqA%2FXXycuWsh08VFRimpNZFKxNS%2BjKGwcelY1MrUbRA3Td%2B1&X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Date=20240123T082939Z&X-Amz-SignedHeaders=host&X-Amz-Expires=300&X-Amz-Credential=ASIAQ3PHCVTY52OJZOFU%2F20240123%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Signature=fac764f4480907c6aaa611d3f0c83200cb9fb45e0cd8f34ad9c3eed29920ab0b&hash=b3d5aed4d5ad3b598be414746fc58771f98226bed5e11f5621477029d50d52c7&host=68042c943591013ac2b2430a89b270f6af2c76d8dfd086a07176afe7c76c2c61&pii=S0275531922001325&tid=pdf-17733653-08a2-4f6d-ae59-d704ce2797a6&sid=e4f6a01418a7b6474b5a54f7a7f286e4a1cfgxrqa&type=cl)

chống SVM: [tại đây](https://www.irjmets.com/uploadedfiles/paper/issue_12_december_2023/47811/final/fin_irjmets1703823161.pdf)