

ỨNG DỤNG THUẬT TOÁN RANDOM FOREST ĐỂ NHẬN DIỆN GIAN LẬN TRONG BÁO CÁO TÀI CHÍNH: BẰNG CHỨNG THỰC NGHIỆM Ở VIỆT NAM

● NGUYỄN THỊ BÍCH THUẬN

Khoa Kinh tế - Luật, Trường Đại học Đồng Tháp

Email: ntbthuan@dthu.edu.vn

TÓM TẮT:

Nghiên cứu được thực hiện nhằm nhận diện gian lận báo cáo tài chính tại các doanh nghiệp niêm yết trên Sở Giao dịch Chứng khoán Thành phố Hồ Chí Minh thông qua nền tảng Tam giác gian lận và thuật toán Random Forest. Kết quả nghiên cứu cho thấy yếu tố Động cơ (ROS và ROA); yếu tố Cơ hội (biến động khoản phải thu và hàng tồn kho) là các biến số quan trọng nhất giúp nhận diện gian lận BCTC. Mô hình này có khả năng dự báo chính xác trên 88% dựa trên chỉ báo AUC-ROC.

Từ khóa: gian lận báo cáo tài chính, Random forest.

1. Đặt vấn đề

Báo cáo tài chính (BCTC) là cầu nối thông tin cốt lõi giữa doanh nghiệp và các bên liên quan (nhà đầu tư, chủ nợ, cơ quan quản lý). Tính trung thực và hợp lý của BCTC là nền tảng duy trì sự minh bạch và hiệu quả vận hành của thị trường vốn. Tuy nhiên, hành vi gian lận BCTC (Financial Statement Fraud - FSF) đang ngày càng tinh vi và gây hậu quả tàn khốc. Theo báo cáo toàn cầu của ACFE (2024), dù chỉ chiếm 5% số ca gian lận doanh nghiệp, FSF lại gây tổn thất nặng nề nhất với mức thiệt hại trung vị lên tới 4,8 triệu USD/vụ. Tại Việt Nam, các đại án thao túng và làm giả số liệu gần đây (như FLC Faros, Tân Hoàng Minh, Vạn Thịnh Phát) không chỉ gây thiệt hại hàng trăm nghìn tỷ đồng cho nhà đầu tư mà còn làm suy giảm chỉ số tín nhiệm, cản trở tiến trình nâng hạng thị trường của chứng khoán Việt Nam (Nguyen & Truong, 2022).

Do đó, việc nhận diện gian lận dựa vào các mô hình thống kê tuyến tính kinh điển như mô hình Z-Score, F-Score, M-Score hay hồi quy Logistic. Tuy nhiên, cấu trúc dữ liệu tài chính hiện đại chứa đựng các mối quan hệ phi tuyến phức tạp và có độ nhiễu cao (Dechow và cộng sự, 2011). Các mô hình cũ thường bỏ sót các tương tác đa chiều giữa các dữ liệu, dẫn đến tỷ lệ bỏ sót gian lận cao. Sự trỗi dậy của khoa học dữ liệu, đặc biệt là thuật toán Random Forest Classification đã giải quyết triệt để bài toán này. Random Forest giúp kiểm soát hiện tượng Overfitting và xử lý xuất sắc bài toán dữ liệu mất cân bằng khi số lượng doanh nghiệp gian lận luôn chiếm tỷ lệ thiểu số (Breiman, 2001; Craja và cộng sự, 2020). Đặc biệt, tính năng Feature Importance của mô hình cung cấp "tính năng dấu hiệu nào giải thích được gian lận ở mức cao, giúp kiểm toán viên định vị chính xác các dấu hiệu cảnh báo trọng yếu dẫn đến gian lận trong báo cáo tài chính.

Chính vì lý do đó, trong bài viết này chúng tôi tiến hành nghiên cứu “Ứng dụng thuật toán Random Forest để nhận diện gian lận trong báo cáo tài chính của các doanh nghiệp niêm yết ở Việt Nam” nhằm cung cấp một giải pháp công nghệ định lượng mạnh mẽ, vừa đóng góp vào lý thuyết kiểm toán hiện đại, vừa mang giá trị ứng dụng thực tiễn cao giúp lành mạnh hóa thị trường vốn Việt Nam.

2. Cơ sở lý thuyết và bằng chứng thực nghiệm

2.1. Khái niệm về gian lận báo cáo tài chính.

Gian lận có thể được định nghĩa theo nhiều cách khác nhau. Trong môi trường kinh doanh, Gian lận báo cáo tài chính là một trong những hành vi gian lận tài chính, nó được xem là các hành vi cố ý và bất hợp pháp dẫn đến các báo cáo bị gây hiểu lầm hoặc tiết lộ tài chính gây hiểu lầm (Beasley 1996) và các bên liên quan bị ảnh hưởng xấu bởi các báo cáo tài chính gây hiểu lầm này (Elliot và Willingham 1980).

Ủy ban Tài trợ của Hoa Kỳ thuộc Ủy ban Treadway (COSO) (Beasley và cộng sự, 1999) và chuẩn mực SAS số 99 (2002) đã chỉ ra một khái niệm gian lận báo cáo tài chính là hành vi cố ý (intentional) hoặc liều lĩnh (reckless conduct) dựa trên thông tin sai lệch hoặc thiếu sót dẫn đến sai lệch đáng kể về báo cáo tài chính.

Ở Việt Nam, theo VSA 240, đoạn 11, ban hành kèm theo Thông tư số 214/2012/TT-BTC, đã xác định: “Gian lận là những hành vi cố ý làm sai lệch thông tin kinh tế, tài chính do một hay nhiều người trong hội đồng quản trị, ban giám đốc, nhân viên hoặc bên thứ ba thực hiện, gây ảnh hưởng đến sự trung thực trên BCTC”.

2.2. Lý thuyết nền giải thích về hành vi gian lận báo cáo tài chính

Có thể nói rằng người đặt nền móng cho nghiên cứu về gian lận báo cáo tài chính là Cressey (1953), người đã phát hiện ra lý thuyết tam giác gian lận. Theo (Cressey, 1953), hành vi gian lận chỉ phát sinh khi hội đủ 3 yếu tố như sau:

- *Áp lực*: là nguyên nhân đầu tiên gây ra gian lận vì áp lực thường xoay quanh tình hình tài chính ở cấp độ cá nhân hoặc công ty (Schuchter & Levi, 2016). Áp lực tài chính buộc các nhà quản lý phải phóng đại doanh thu để tăng hiệu quả hoạt động của công ty một cách giả tạo nhằm duy trì hoạt động và hưởng các chính sách khen thưởng (Dorminey, et al., 2010). Các

đại diện của áp lực bên ngoài đề cập đến nhu cầu tài trợ liên quan đến việc tạo ra tiền mặt nội bộ (Dechow và cộng sự 1996) được điều chỉnh bằng cổ tức bằng tiền mặt và chi tiêu vốn.

- *Cơ hội thực hiện gian lận*: Việc nắm rõ các cơ chế kiểm soát nội bộ và khả năng phát hiện đóng vai trò quan trọng trong việc ảnh hưởng đến khả năng cá nhân thực hiện gian lận (Ramamoorti, et al., 2013; Rezaee & Riley, 2009).

- *Sự hợp lý hóa*: Những người vi phạm thường không coi mình là tội phạm và biện minh cho hành động của họ bằng các yếu tố bên ngoài.

2.3. Các nghiên cứu thực nghiệm liên quan

Nhiều nghiên cứu gần đây đã sử dụng kỹ thuật học máy kết hợp, đặc biệt là kỹ thuật Random Forest classifiers để xác định các doanh nghiệp có dấu hiệu gian lận báo cáo tài chính và khẳng định rằng mô hình này có ý nghĩa và đáng tin cậy chẳng hạn:

Patel và cộng sự (2018) thực hiện nghiên cứu nhận diện gian lận BCTC bằng thuật toán Random Forest trên tập dữ liệu 86 gian lận BCTC và 92 không gian lận BCTC của các công ty sản xuất niêm yết ở Sở giao dịch chứng khoán Bombay trong giai đoạn 2008–2011. Kết quả nghiên cứu cho thấy có 10 biến tỷ số tài chính dự báo tốt cho gian lận BCTC bao gồm các yếu tố như ROA, ROS, thu nhập mỗi cổ phần, giá trị cổ phiếu; các yếu tố đòn bẩy tài chính, vòng quay khoản phải trả, và các yếu tố thuộc vốn lưu động. Sau đó, Sử dụng 42 thuật toán khác nhau kết quả cho rằng thuật toán random Forest cho độ chính xác được cải thiện tốt nhất đạt 81,4%.

Nghiên cứu của Ye và cộng sự (2019) sử dụng dữ liệu thử nghiệm bao gồm 11726 báo cáo tài chính công khai của Trung Quốc từ năm 2007 đến năm 2017, trong đó 1314 báo cáo tài chính bị Ủy ban Điều tiết Chứng khoán Trung Quốc (CSRC) cáo buộc gian lận. Nghiên cứu sử dụng nhiều thuật toán khác nhau như: Random Forest (RF), Mạng thần kinh nhân tạo (ANN), Hồi quy logistic (LR), Máy vector hỗ trợ (SVM), CART, , Mạng Bayesian, Bagging, Stacking và Adaboost. Kết quả cho thấy thuật toán Random Forest cho kết quả vượt trội nhất với độ chính xác đạt trung bình 86,9%, trong khi các thuật toán ANN và LR có độ chính xác trung bình lần lượt là 73% và 64,5%.

Nghiên cứu Nien và cộng sự (2024) sử dụng các chỉ số tài chính và phi tài chính làm thước đo để phát

hiện gian lận BCTC. Với dữ liệu bảng gồm 2235 quan sát của các công ty niêm yết trên Sở giao dịch chứng khoán Việt Nam từ năm 2014 đến năm 2020. Dựa trên nguyên tắc trọng yếu trong kiểm toán, nghiên cứu đã chia tỷ lệ chênh lệch lợi nhuận thành bốn ngưỡng gian lận trọng yếu là trên 5%, 10%, 20% và 50%. Hai kỹ thuật khai thác dữ liệu đã được sử dụng là Random Forest và ANN để xây dựng mô hình dự đoán gian lận tốt nhất. Kết quả cho thấy độ chính xác trung bình của kết quả dự đoán của thuật toán RF đạt 91% đối với ngưỡng gian lận trọng yếu là 5%; khi mức độ trọng yếu của gian lận tăng lên trên 50% chênh lệch lợi nhuận, khả năng dự đoán đạt 98%. Mặt khác, thuật toán RF xác định được các yếu tố tài chính có khả năng phát hiện gian lận BCTC sớm bao gồm ROA, ROE và lợi nhuận trước thuế trên tổng tài sản.

Nghiên cứu gần đây nhất của Nguyễn Tiến Hùng và Võ Hồng Đức (2017) hướng tới mục tiêu xây một mô hình trong việc phát hiện gian lận BCTC của các DNNY [1]. Các tác giả đã tiến hành nghiên cứu dựa

trên BCTC của 88 doanh nghiệp niêm yết trên HOSE (44 doanh nghiệp có gian lận và 44 doanh nghiệp không có gian lận BCTC). Kết quả nghiên cứu cho thấy mô hình có khả năng dự báo chính xác đến 68% các doanh nghiệp thuộc mẫu nghiên cứu (dự báo chính xác đến 75% đối với các doanh nghiệp có gian lận và 61% đối với các doanh nghiệp không gian lận).

3. Dữ liệu, phương pháp và mô hình nghiên cứu

3.1. Mô hình nghiên cứu

Dựa trên lý thuyết tam giác gian lận (Cressey, 1953) và các nghiên cứu thực nghiệm ở trên bài viết xây dựng mô hình nghiên cứu như sau

$$FRAUR_{it} = \beta_0 + \beta_1 \sum_{it} AL_{it} + \beta_2 \sum_{it} CH_{it} + \beta_3 \sum_{it} HLH_{it} + \varepsilon_{it}$$

Trong đó:

$\sum_i AL_i$ là tập hợp các biến thể hiện yếu tố áp lực.

$\sum_i CH_i$ là tập hợp các biến thể hiện yếu tố cơ hội.

$\sum_i HLH_i$ là tập hợp các biến thể hiện yếu tố hợp lý hóa.

Bảng 1. Tóm tắt đo lường các biến số trong mô hình nghiên cứu

Nhân tố	Biến	Đo lường Phương pháp
Biến phụ thuộc		
FRAUD		0: Bình thường 1: Doanh nghiệp có dấu hiệu gian lận hoặc bị xóa khỏi danh sách niêm yết vì lý do liên quan đến sự không minh bạch trong BCTC .
Biến độc lập		
Áp lực	LEV	Tổng nợ/Tổng tài sản
	ROA	Lợi nhuận ròng/Tổng tài sản
	CF	Logarit tự nhiên của dòng tiền tự do
	ROS	Lợi nhuận ròng/Tổng tài sản
Cơ hội	AR	Tài khoản phải thu (hiện tại) năm) – Tài khoản phải thu (trước đó) năm)
	AI	Hàng tồn kho (Năm hiện tại – Hàng tồn kho (năm trước đó)
	SIZE	Logarit tự nhiên của Tổng Tài sản
Hợp lý hóa	Audit	0: BCTC của công ty được kiểm toán bởi 4 công ty kiểm toán uy tín (Big 4). 1: BCTC của công ty không được kiểm toán bởi 4 công ty kiểm toán uy tín (Big 4). 2: BCTC của công ty không được kiểm toán bởi 4 công ty kiểm toán uy tín (Big 4). Và đã có thay đổi công ty kiểm toán qua nhiều năm liên tiếp . 3: Công ty đã thay đổi kiểm toán, trước đó công ty kiểm toán là Big4 4: BCTC của công ty chưa được công ty nào kiểm toán.

3.2. Dữ liệu nghiên cứu

Dữ liệu được thu thập từ BCTC của các công ty phi tài chính niêm yết trên Sở giao dịch chứng khoán Thành phố Hồ Chí Minh (HSX) và Hà Nội (HNX) từ năm 2018 đến năm 2022. Các công ty tài chính, chẳng hạn như ngân hàng, công ty bảo hiểm và công ty chứng khoán, không được xem xét trong nghiên cứu này vì đặc điểm báo cáo tài chính của họ khác biệt đáng kể so với các công ty phi tài chính. Các tổ chức tài chính hoạt động theo các khuôn khổ pháp lý và chuẩn mực kế toán riêng biệt do bản chất mô hình kinh doanh của họ, dựa trên tài sản, nợ phải trả và yêu cầu vốn điều lệ. Những khác biệt này có thể ảnh hưởng đến khả năng áp dụng của các mô hình trong việc phát hiện gian lận trong báo cáo tài chính. Bằng cách tập trung vào các công ty phi tài chính, chúng tôi

Random Forest là thuật toán học máy dựa trên tập hợp nhiều cây quyết định, giúp nâng cao độ chính xác dự báo và giảm hiện tượng overfitting. Mô hình được huấn luyện trên tập dữ liệu doanh nghiệp với biến phụ thuộc là trạng thái gian lận (fraud = 1 nếu doanh nghiệp có gian lận; fraud = 0 nếu ngược lại). Hiệu quả mô hình được đánh giá thông qua Accuracy, Precision, Recall, F1-score và ROC-AUC. Đồng thời, nghiên cứu sử dụng Feature Importance để xác định các yếu tố tài chính quan trọng trong việc dự báo gian lận báo cáo tài chính. RF được coi là một phương pháp chính xác và mạnh mẽ vì số lượng cây quyết định tham gia vào quá trình này. Nó không bị vấn đề quá khớp. Lý do chính là nó lấy trung bình của tất cả các dự đoán, giúp loại bỏ các sai lệch.

4. Kết quả nghiên cứu

Bảng 2. Kết quả phân tích mức độ ảnh hưởng các biến số trong mô hình nghiên cứu

	Tên biến							
	Áp lực				Cơ hội			Hợp lý hóa
	ROA	LEV	CF	ROS	AR	AI	SIZE	Audit
Feature Importance	33.18%	5.63%	3.73%	10.71%	21.85%	13.35%	6.80%	1.00%

đảm bảo tính nhất quán và phù hợp của các phương pháp phát hiện gian lận được áp dụng, chẳng hạn như sử dụng các kỹ thuật học máy, phù hợp hơn với các công ty có cấu trúc doanh thu, chi phí và tài sản truyền thống.

3.3. Phương pháp nghiên cứu

Nghiên cứu sử dụng thuật toán Random Forest, cùng với phần mềm google colab để dự báo khả năng gian lận báo cáo tài chính của doanh nghiệp.

Thuật toán Radom Forest (RF) được tạo ra bởi Ho (1995). Một phần mở rộng của thuật toán đã được phát triển bởi Breiman (2001). RF hoạt động theo bốn bước:

- Bước 1: Chọn một mẫu ngẫu nhiên từ tập dữ liệu đã cho;
- Bước 2: Thiết lập một cây quyết định cho mỗi mẫu và nhận được kết quả dự đoán từ mỗi cây quyết định;
- Bước 3: Bỏ phiếu cho mỗi kết quả dự đoán;
- Bước 4: Chọn kết quả được dự đoán nhiều nhất làm dự đoán cuối.

Kết quả mô hình Random Forest cung cấp những bằng chứng quan trọng về mức độ tác động của các biến đến khả năng nhận diện gian lận BCTC. Đáng chú ý, các biến số đại diện cho "Áp lực tài chính" gồm ROA (33,18%) và ROS (10,31%) nắm giữ tỷ trọng đóng góp lớn nhất vào sức mạnh dự báo của mô hình. Điều này chứng minh áp lực sụt giảm khả năng sinh lời và biên lợi nhuận là động cơ trực tiếp buộc ban điều hành phải can thiệp vào số liệu để làm đẹp bức tranh kinh doanh trước thị trường vốn. Kết quả này phù hợp với lý thuyết nền và các nghiên cứu thực nghiệm trước đây.

Bên cạnh đó, Các biến số đại diện cho yếu tố "Cơ hội" gồm biến động của khoản phải thu (21,85%) và hàng tồn kho (13,35%) có trọng số rất cao vào giải thích mức độ dự báo gian lận BCTC của mô hình. Phát hiện này hoàn toàn đồng thuận với lý thuyết tam giác gian lận và nghiên cứu thực nghiệm của Skousen và cộng sự (2009), khẳng định ban quản lý đã lạm dụng các "vùng xám" chủ quan trong chu kỳ phải thu và tồn kho để thực hiện các thủ thuật dồn tích nhằm thao túng dữ liệu BCTC.

Bảng 3. Kết quả mức độ chính xác của việc sử dụng thuật toán Random Forest

	Accuracy	Precision	Recall	AUC-ROC	Time (s)
Random Forest	77.4%	21.0%	85.0%	88.6%	420

Các yếu tố khác còn lại trong mô hình có trọng số không đáng kể trong việc giúp mô hình phân loại doanh nghiệp gian lận và không gian lận.

Kết quả mô hình cho thấy

- Accuracy = 77,4% chứng tỏ mô hình RF đúng dấu hiệu gian lận BCTC khoảng 76,26% tổng số doanh nghiệp trong mẫu nghiên cứu. Vì dữ liệu mất cân bằng nên cần xem xét mức độ chính xác của dự báo thông qua AUC-ROC và Recall.

- AUC-ROC đạt 88,4%: Đạt ngưỡng rất tốt theo tiêu chuẩn học máy tài chính (Craja và cộng sự, 2020). Chỉ số này chứng minh rằng mô hình Random Forest sở hữu năng lực phân tách phân phối xác suất rủi ro giữa doanh nghiệp gian lận và doanh nghiệp minh bạch một cách cực kỳ mạnh mẽ, bất chấp độ nhiễu của dữ liệu cận biên.

- Độ nhạy đối với gian lận (Recall Class 1 = 85%): Mô hình nhận diện chính xác 62 trên tổng số 73 vụ gian lận thực tế, chỉ bỏ sót 11 trường hợp. Tỷ lệ Recall đạt 85% khẳng định giá trị ứng dụng thực tiễn cao của mô hình trong vai trò một công cụ sàng lọc và cảnh báo sớm cho các cơ quan quản lý và các định chế tài chính.

- Mặc dù tỷ lệ chính xác trong phân loại gian lận

(Precision Class 1) chỉ đạt 21%, điều này hoàn toàn phù hợp với bản chất của các mô hình với dữ liệu không cân bằng và không làm giảm đi giá trị thực tiễn của nghiên cứu.

5. Kết luận và hàm ý chính sách

Nghiên cứu đã ứng dụng thành công mô hình học máy tiên tiến Random Forest Classification để giải quyết bài toán phát hiện gian lận báo cáo tài chính trong bối cảnh dữ liệu mất cân bằng nghiêm trọng. Với chỉ số AUC-ROC đạt 88,4% và Recall đạt 85%, mô hình vượt trội hoàn toàn so với các phương pháp thống kê tuyến tính truyền thống trong việc giảm thiểu sai lầm loại II (bỏ sót gian lận).

Nghiên cứu chứng minh mối liên kết chặt chẽ giữa các thuộc tính toán học của học máy và các cấu phần lý thuyết nền kinh điển: Áp lực tài chính (đo bằng ROA) và Cơ hội từ đặc điểm tài sản (đo bằng biến động khoản phải thu và biến động hàng tồn kho) là hai yếu tố cốt lõi chi phối hành vi gian lận BCTC. Kết quả này cung cấp một giải pháp định lượng mạnh mẽ cho các kiểm toán viên, các nhà đầu tư và các cơ quan quản lý thị trường vốn trong việc chủ động nhận diện và ngăn chặn các rủi ro hệ thống, hướng tới sự minh bạch và bền vững của nền kinh tế ■

TÀI LIỆU THAM KHẢO:

Bộ Tài chính Việt Nam (2012). Chuẩn mực kiểm toán Việt Nam số 240 - Trách nhiệm của kiểm toán viên đối với gian lận trong kiểm toán báo cáo tài chính.

Nguyễn Anh Phong. Chuyên đề: Kỹ thuật hiện đại (AI) nhận diện gian lận BCTC doanh nghiệp: Tổng quan lý thuyết. Trường Đại học Kinh tế - Luật, Đại học Quốc gia TP.HCM.

Nguyễn Tiến Hùng, Võ Hồng Đức, (2017). Nhận diện gian lận báo cáo tài chính: Bằng chứng thực nghiệm tại các doanh nghiệp niêm yết ở Việt Nam, Tạp chí Công nghệ Ngân hàng, 132 (2017) 5, 58-72.

Beasley, M. S. (1996). An empirical analysis of the relation between the board of director composition and financial statement fraud. *Accounting Review*, 443-465.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.

Cressey, D. R. (1950). The Criminal Violation of Financial Trust. *American Sociological Review*, 15(6), 738-743.

Chen, F. H., Chi, D. J., & Zhu, J. Y. (2014, August). Application of random forest, rough set theory, decision tree and neural network to detect financial statement fraud-taking corporate governance into consideration. In *International conference on intelligent computing* (pp. 221-234). Cham: Springer International Publishing.

- Dechow, P. M., Ge, W., Larson, C. R., & Sloan, R. G. (2011). Predicting material accounting misstatements. *Contemporary accounting research*, 28(1), 17-82.
- Nhien, C. T., Hung, D. N., & Bình, V. T. T. (2024). Using random forest and artificial neural network to detect fraudulent financial reporting: Data from listed companies in Vietnam. *Calitatea*, 25(202), 160-173.
- Patel, H., Parikh, S., Patel, A., & Parikh, A. (2018). An application of ensemble random forest classifier for detecting financial statement manipulation of Indian listed companies. In *Recent Developments in Machine Learning and Data Analytics: IC3 2018* (pp. 349-360). Singapore: Springer Singapore.
- Skousen, C. J., Smith, K. R., & Wright, C. J. (2009). Detecting and predicting financial statement fraud: The effectiveness of the fraud triangle and SAS No. 99.
- Trang, N. T. N., Thuan, L. T., Nhi, N. D. L., & Tu, V. C. (2025). Unravelling Financial Statement Fraud in Emerging Markets: Insights from Vietnam with the Fraud Diamond Framework. *International Journal of Applied Economics, Finance and Accounting*, 22(1), 68-81.
- Ye, H., Xiang, L., & Gan, Y. (2019). Detecting financial statement fraud using random forest with SMOTE. In *IOP conference series: materials science and engineering* (Vol. 612, No. 5, p. 052051). IOP Publishing. <https://doi.org/10.1287/mnsc.1050.0487>

Ngày nhận bài: 2/3/2026

Ngày phản biện đánh giá và sửa chữa: 18/3/2026

Ngày chấp nhận đăng bài: 3/4/2026

APPLICATION OF THE RANDOM FOREST ALGORITHM TO DETECT FRAUD IN FINANCIAL STATEMENTS: EMPIRICAL EVIDENCE FROM VIETNAM

● **NGUYEN THI BICH THUAN**

Faculty of Economics and Law, Dong Thap University

Email: ntbtthuan@dthu.edu.vn

ABSTRACT:

This study aims to detect financial statement fraud among listed companies on the Ho Chi Minh City Stock Exchange using the fraud triangle framework and the Random Forest algorithm. The study results show that motivation factors, including ROS and ROA, and opportunity factors, including changes in receivables and inventories, are the most important variables for identifying financial statement fraud. The model is capable of achieving prediction accuracy of over 88% based on the AUC-ROC indicator.

Keywords: fraud, financial statements, algorithm, Random Forest.