Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 99–109
DOI: 10.15439/2025F9424
ISSN 2300-5963 ACSIS, Vol. 43

CADM: An LSTM-Based Model for Detecting Creative Accounting in Time-Series Data from Saudi-Listed Companies

Maysoon Bineid
0009-0007-4447-3706
Department of IS and Technology University of Jeddah, SA Department of Informatics
University of Sussex, Brighton. Falmer, BN1 9RH, UK
Email: M.bineid@sussex.ac.uk

Natalia Beloff
0000-0002-8872-7786
Department of Informatics, University of Sussex, Brighton. Falmer, BN1 9RH, UK
Email: N.Beloff@sussex.ac.uk

Anastasia Khanina
0009-0007-7496-0054
School of Business and Law, University of Brighton
Brighton, BN2 4NU, UK
Email: A.Khanina@brighton.ac.uk

Martin White
0000-0001-8686-2274
Department of Informatics, University of Sussex, Brighton. Falmer, BN1 9RH, UK
Email: M.White@sussex.ac.uk

***Abstract*—Studies on Saudi accounting practices have identified evidence of creative accounting in the financial statements of listed companies. Despite the application of various fraud detection methods, identifying legal but misleading manipulations remains challenging. This paper extends the Creative Accounting Detection Model (CADM), an LSTM-based model originally proposed by Bineid et al. (2023, 2024) for detecting creative accounting. Two versions, (CADM1) and (CADM2), were trained on two simulated datasets with different bases, achieving 100% and 95% accuracy, respectively. Testing on the energy sector (2019-2023), CADM1 identified one company as engaging in creative accounting, while CADM2 classified all companies as non-creative with greater confidence stability. The findings establish CADM as a robust, scalable solution for the early detection of financial manipulation. By combining predictive strength with explainability, CADM can be employed to advance current approaches to forensic accounting and risk analytics, offering valuable insights to regulators, auditors, and decision-makers.**

***Index Terms*—Deep Learning, Long Short-Term Memory, LSTM, Creative Accounting, Saudi Arabia.**

I. Introduction

DESPITE the establishment of accurate accounting standards, financial statement scandals continue to shake global markets. Cases such as Enron [1] and Wirecard [2] reveal the persistent challenge of Creative Accounting (CA), a practice where managers manipulate the accounting figures to present a misleading image of financial position.

Business performance is typically evaluated through accounting outcomes presented in financial statements (FSs). To ensure these statements offer a true and fair view, companies are required to prepare them under well-defined accounting standards and to subject them to independent auditing [3]. These measures are intended to ensure transparency, confirm compliance, and detect any material misstatement. However, corporate managers often have incentives to present the most favourable image of their organisation [4]. In doing so, they may engage in CA practices, adjusting figures within the boundaries of acceptable standards, without technically violating the rules. Although CA is sometimes distinguished from fraud, its consequences can be just as damaging [5], [6].

While many studies have focused on detecting Financial Statement Fraud (FSF), fewer studies have tried to identify CA (also referred to as Earnings Management EM) [7], [8]. One of the earliest and most notable efforts is the Beneish M-Score [9], a model built on financial ratios designed to flag potential EM. However, its capability in a non-U.S. context remains questionable [10]. Other statistical models like the Jones model and the modified Jones model [11] have been developed to detect EM, but they still face limitations such as model misspecification and low detection power, particularly in varying economic contexts.

More recently, Deep Learning (DL) applications have emerged as a powerful tool in financial prediction and classification. DL models, particularly neural networks, have demonstrated promise in learning complex patterns embedded in large unstructured datasets. Approaches combining Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), Gated Recurrent Network (GRN), and even Bidirectional Encoder Representations from Transformers (BERT) models have been explored to solve problems like gradient vanishing and square modelling [12], [13]. For example, Craja et al. [14] applied deep learning techniques to detect FSF based on textual analysis of financial reports using a Bag-of-Words (BOW) approach. The model successfully identified fraudulent cases; however, it was designed for linguistic analysis only. Schreyer et al. [15] used Adversarial Autoencoder Neural Networks to detect anomalies in raw accounting entries, although their focus on ERP data structures limits the ability to generalise to standardised FSs.

Given the sequential nature of FSs, LSTM networks have shown remarkable suitability for analysing time-series data. They are particularly adept at capturing long-term dependencies and detecting subtle deviations from normal patterns. Their success in areas like anomaly detection, image recognition, and natural language processing has encouraged their application to financial domains, including fraud detection [16], [17] and performance predictions [18]. However, applying DL techniques to CA detection requires

Topical area: Advanced Artificial Intelligence in Applications

overcoming a fundamental obstacle: the lack of real-world CA examples. However, in the Saudi Arabian context, studies have found that FSs do not represent a company's true and fair position, despite auditor procedures approval[4], [19], [20], [21]. This alarming finding emphasises the need for innovative solutions to detect FS manipulation and improve financial reporting quality.

Despite the growing literature on detecting financial statement fraud, a significant gap remains in detecting CA weather through statistical methods or Artificial Intelligence, particularly within the Saudi Arabian context. In addition, the absence of labelled datasets for CA, particularly in the Saudi market, presents a methodological challenge in developing, training, and validating such models. To address this gap, this research aims to design, simulate, and evaluate a Creative Accounting Detection model (CADM), using an LSTM architecture, capable of learning from domain-specific financial and non-financial patterns.

To achieve this aim, the study makes several contributions to the literature. First, it introduces a novel domain-specific simulation methodology tailored to the characteristics of Saudi-listed companies. Second, it compares two distinct simulation approaches to determine which offers more effective training for the CA detection model. Third, it validates the proposed models using real-world FS data from the Saudi energy sector. These contributions are guided by the following research questions:

- **RQ1:** How does the base used in simulated data affect the accuracy of LSTM models in detecting creative accounting?
- **RQ2:** Can the proposed deep learning model (CADM) be trained to classify with high accuracy?
- **RQ3:** Can the proposed deep learning model (CADM) be generalised to real-world FSs?

II. Methodology

A mixed-methods approach is employed to deliver the research objectives. It combines an experimental methodology with quantitative techniques to form a framework for developing, training, and evaluating a deep-learning model using simulated and real-world datasets. The upcoming subsections explain the research design, feature engineering, data preparation, model architecture, training process, and testing procedures.

A. Research Design

The research design comprises two overlapping phases: quantitative and experimental, implemented in parallel across different stages of this study. The first phase involves preparing the dataset, which consists of two main components: a training dataset and a testing dataset. The testing dataset is collected from real-world FSs of Saudi-listed companies. However, due to the unavailability of a CA-labelled dataset, this study adopts a simulation-based methodology to generate FSs that reflect selected CA patterns. This leads us to start the second phase before ending the first, as shown in Fig. 1.

This part of the experimental phase involves implementing a domain-specific simulation process to generate synthetic datasets. However, two simulation designs were applied simultaneously. The first simulation is based on the real-world dataset analysis, and the second is based on the findings of Leitch and Chen [21].

Once the simulation stage is completed, the quantitative phase resumes with the development and training of the model. Using an LSTM-based architecture, the quantitative approach is implemented by analysing numeric accounting data, financial ratios, Corporate Governance (CGs), and key performance indicators extracted from real and simulated FSs to capture and classify CA practices. The outcomes of this analysis are further evaluated using deep learning performance metrics in the remainder of the second phase.

B. Data Sources and Collection

The emergence of publicly accessible financial data in Saudi Arabia is a relatively recent advancement that reflects the Kingdom's efforts to transform its financial system, in-

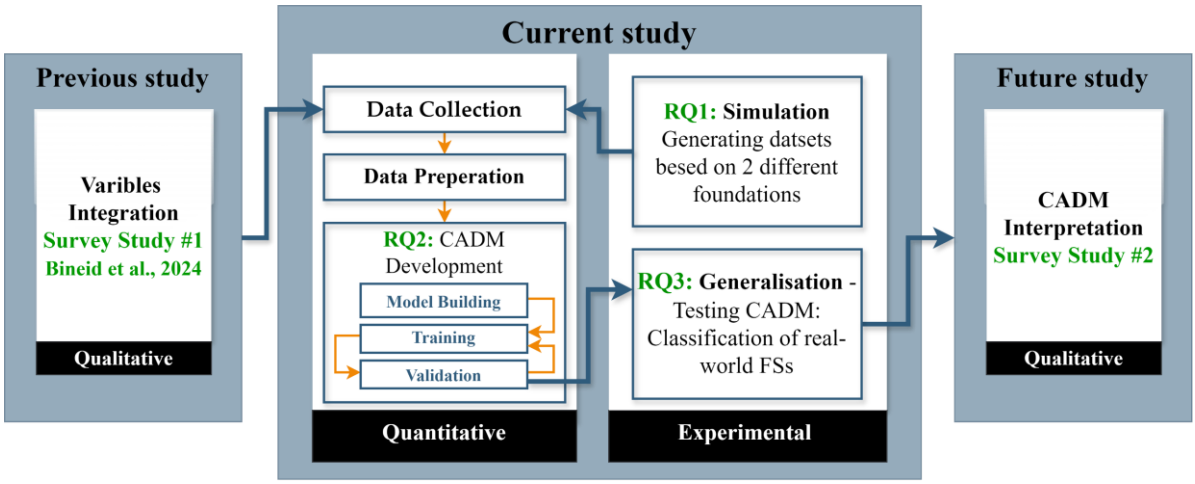

Fig. 1: CADM Research Framework

crease transparency, and attract domestic and foreign investment. These efforts are rooted in institutional reforms that began in the early 2000s, particularly with the creation of the CMA and have accelerated under the Vision 2030 economic transformation plan[22].

Saudi Arabia's enforcement of financial disclosure standards and IFRS compliance since 2017 has significantly improved the reliability and usability of its financial datasets. Real-world financial data can be accessed from the Saudi Exchange online platform (Tadawul), which provides official financial disclosures, including downloadable PDFs and Excel data. CMA and SOCPA platforms also offer regulatory filings, enforcement actions, historical announcements, standards, conversion guidelines, and auditor regulations. Third-party aggregators, such as Argaam.com and Mubasher.com, offer processed versions of the raw data; nevertheless, these platforms are commercially operated entities, not government-affiliated sources.

The Saudi market has 22 different sectors, each representing a distinct market segment with a varying number of companies in each sector. The largest sector by market capitalisation is the Energy sector, and the smallest is the Entertainment which has a limited number of companies with a relatively lower market capitalisation compared to others. Sectors and the exact number of companies can vary due to new listings and reclassifications in the market. However, FSs of Saudi-listed companies are publicly available in several formats. Tadawul provides financial data for all stakeholders, available in several formats, either online or offline. They also provide

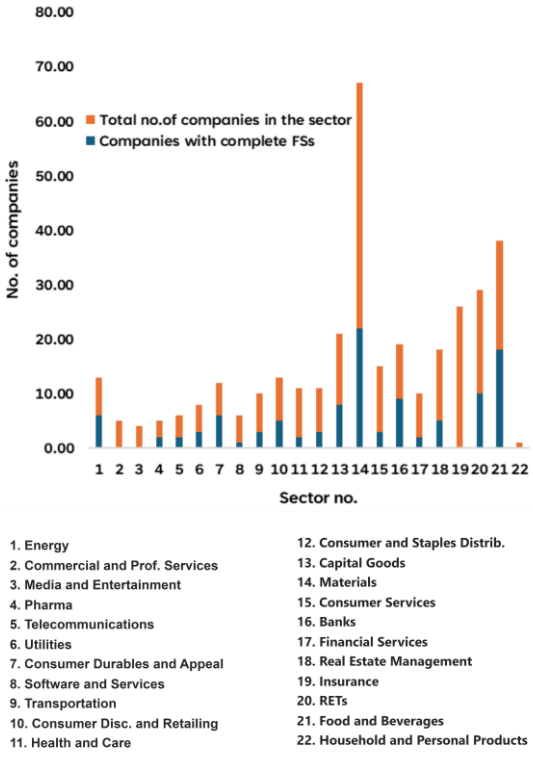

Fig. 2: Availability of FSs in the Saudi Market

a useful analytical tool for paid subscription users. But still, some companies fail to publish regularly in Tadawul, as shown in Fig. 2.

The Energy sector, having 7 companies, was chosen for this study. Although the Materials sector has the highest number of companies and is academically recognised for its high consistency across firms, making data preparation relatively easier, there was a concern that such uniformity might restrict the model's ability to learn diverse patterns and ultimately reduce training effectiveness. To better challenge the model and assess its generalisation capabilities, the Energy sector was selected due to the greater diversity among its companies.

C. Simulation Strategy

Although there is an increase in data availability, real-world examples of CA are challenging to find, particularly in the Saudi context [7], as companies do not usually admit publicly to using CA techniques. Without a real dataset, the best alternative would be to employ proxies, such as accrual-based analysis [26], on publicly available financial datasets to infer potential earnings management. However, accrual-based models like the Beneish M-Score [9] rely on assumptions and may not provide a definitive ground truth for CA, as they may include false positives and negatives, admitted by Beneish himself [10].

Moreover, AI strategies that can effectively address the lack of data in training deep learning models, such as Semi-supervised Learning, cannot be used because they lack the level of customisation needed in the dataset, and their results will be difficult to generalise [23]. It can also be helpful to use pre-trained models on a broader dataset (e.g., fraud detection models) as in Transfer Learning. Still, they are more likely to perform poorly on the domain-specific (Saudi-listed companies), and controlled variables (selected CA patterns) used in this research.

Consequently, simulating financial data appeared to be the optimal strategy that fulfils the need to have a sector-specific dataset with labelled data and controlled variables. Simulating financial statements is a well-established practice in the literature, employed for various analytical purposes. For instance, Leitch and Chen [24] simulated monthly FSs to evaluate key financial indicators and explore broader organisational dynamics.

Building upon their methodology, this study simulates two datasets: one based on general accounting measures and another tailored to the characteristics of the Saudi-listed companies, including financial (FIN) and non-financial (N-FIN) variables. Each dataset is prepared to train different versions of the CADM model. Generating a simulated dataset can explicitly encode CA techniques relevant to the Saudi business environment, providing flexibility for training scenarios and control over manipulation patterns, specifically when the base of the simulation is the real-world dataset. In other words, collecting the real-world dataset is essential to start the simulation process.

A structured pattern-based financial simulation approach has been utilised to generate the training dataset. First, two

primary goals were established: (1) to simulate a dataset closely resembling real-world data to ensure the model is effectively trained and (2) to account for four different scenarios of account manipulations, given the lack of clarity about the exact manipulation scenarios present in real datasets, if any. The dataset consists of two groups of companies: labelled as CA and labelled as N-CA (Non-Creative Accounting). The N-CA group was generated without applying any manipulations, while the CA group underwent an additional process where CA patterns were embedded.

To achieve the first goal, FIN features such as raw accounting data and financial ratios, and NFIN features such as CG metrics and auditor's status, were incorporated into the simulation process. To address the second objective of this study, two datasets were simulated using two different bases: one based on general accounting measures reported in the literature and the other grounded in real-world data analysis. The first simulated dataset (DST_DSL) was generated by adopting the simulation methodology proposed by Leitch and Chen [24]. For the second simulated dataset (DST_DSR), initial statistical analyses of the real-world dataset were conducted and used as the baseline for determining measurements and ratios. This dual approach enabled the training of two separate models on distinct simulated datasets, thereby enhancing the robustness and adaptability of the proposed detection framework. A summary of the datasets used in the processes is provided in Table I.

Both simulated datasets are validated before being used in training; they are statistically analysed to compare key financial metrics with the real-world dataset, ensuring that trends, outliers, and energy sector characteristics are aligned. This was performed iteratively; when the validation results were not acceptable, the simulation was redesigned, and the generation phase was restarted.

TABLE I: DATASET NAMING AND SPECIFICATIONS USING MATLAB

Dataset	Dataset type	Base	Purpose
DST_DSL.mat	Simulated	[20]	Training CADM1
DSV_DSL.mat	Simulated	[20]	Validation CADM1
DST_DSR.mat	Simulated	DSR.mat	Training CADM2
DSV_DSR.mat	Simulated	DSR.mat	Validation CADM2
DSR.mat	Real-world	N/A	Testing CADM1&2

D. Feature Selection and Engineering

This study aims to simulate a set of widely recognised and frequently cited CA techniques, as in [25], [26], within a controlled dataset, representing them as manipulation patterns for CADM to learn from and detect. Four key CA patterns were selected for inclusion in the simulated dataset: revenue inflation, underreporting of COGS, assets overstatement, and underreporting of liabilities, as detailed in Table II. To make sure these patterns are relevant to the Saudi context, an additional layer of selection criteria was applied. Specifically, these techniques were also chosen based on their prevalence in the local literature [4], [19], [27], their documentation in real-world regulatory reports [28], and their alignment with the findings of the integration study [7].

To simulate a realistic and varied dataset, not all four CA techniques were applied to all companies in the CA label. Instead, each manipulated company was randomly assigned one of the four techniques, which was then consistently applied across all five-year financial periods. This approach avoids compound distributions that may arise from stacking multiple manipulation techniques and allows the model to learn distinct behavioural patterns associated with each manipulation type. It also better reflects real-world conditions, where companies typically engage in certain earnings management strategies over others based on their internal practices, industry norms, and regulatory pressure.

While other CA techniques have been identified in the literature, the selected CA patterns affect quantifiable financial ratios that make them suitable for detection by deep learning models based on time-series patterns. They are also compatible with the simulation process, as each technique can be implemented in a consistent and replicable way across multiple years. In addition, these patterns evolve, which is a key feature that aligns well with the capabilities of LSTM-based architectures. Finally, it is important to limit the number of manipulation types to control the model's complexity and reduce the risk of introducing noise or overlapping effects that could compromise the model's interpretability and reliability

TABLE II: SIMULATED TECHNIQUES OF CA

CA technique	**Behaviour in FSs**	**Simulation**
Revenue inflation	Revenue overstatement is one of the commonly used earnings management strategies and directly impacts key ratios such as net profit margin, EPS, and ROA [1], [2], [3].	A 15% inflation is applied to the revenue variable for each year for the generated companies.
Underreporting of COGS	This manipulation affects profitability without inflating top-line revenue and is less obvious than revenue inflation, thus posing a greater challenge for detection models[3], [4].	Reported COGS are decreased 10% (manipulated by 0.9) for each year while keeping the revenue constant.
Assets overstatement	This technique can distort key financial ratios such as ROA, current ratio, and debt-to-equity ratio. This tactic is commonly observed in cases where firms attempt to hide their losses or signal strength to creditors and investors [5].	Asset value is increased by 10% for each year across all years.
Liabilities underreporting	This practice is used to present a healthier financial position. It improves leverage ratios (debt-to-equity) and conceals financial risk. This pattern is particularly relevant in the credit evaluation context [3], [6].	Liabilities were reduced by 10%for each year, thereby relatively overstating equity.

Both financial (FIN) and non-financial (NFIN) indicators associated with the selected patterns were incorporated into the model. These indicators are typically variables integrated from financial reports and FSs. The selection of FIN variables was based on their relationship with the accounting pattern under examination. In contrast, non-financial variables were incorporated based on their demonstrated significance in identifying accounting manipulation, as evidenced by the findings of the integration study [7]. The model included 14 FIN (FIN1-FIN14) and 16 NFIN variables (NFIN1-NFIN20) as in Table III.

E. Model Architecture

The next step after data preparation and the identification of CA patterns is to select an appropriate deep learning technique for developing the predictive model CADM. As previously discussed, LSTM neural networks are widely regarded as one of the most effective methods for detecting unexpected anomalies in sequential unstructured data spanning multiple years. As an advanced type of RNN, LSTM is specifically designed to learn long-term dependencies in time series through a gated architecture. These gates, responsible for retaining, discarding, or updating information, enable the model to capture complex temporal relationships in historical financial statements.

Unlike traditional models, LSTM is robust against the need for extensive manual feature engineering, as it can automatically learn features from raw data. This allows for the inclusion of diverse FIN and N-FIN variables. Moreover, while CADM is designed to be trained using FIN and Non-FIN data, LSTM excels at modelling this non-linearity in data, specifically when it is required to capture temporal and contextual variables such as CGS measures, regulations, and multi-year data. This makes it possible to build a powerful deep learning model tailored to the Saudi business context. To effectively leverage these advantages, the CADM was implemented using multiple layers, where each layer contributes to progressively learning higher-level temporal features from the FIN and N-FIN input data. All LSTM-related mathematical formulations used in this study are adapted from Hochreiter and Schmidhuber [29], unless otherwise noted. The layered architecture (illustrated in Table III) is composed as follows:

Input Layer

The input layer receives sequential financial data spanning multiple years, structured as a time-series input as in Fig. 3. Each input sequence represents a company's FIN and N-FIN variables over consecutive reporting periods, allowing the

TABLE III: FIN AND N-FIN VARIABLES

FIN			NFIN			
Variable		Code	Variable		Code	Encoding
Accounting raw data	Bank balance and cash	FIN1	Basic	Year	NFIN1	2019-2023
	Inventory	FIN2		Company name	NFIN2	Table 1 Appendix
	Total assets	FIN3		Auditor Firm	NFIN3	Table 2 Appendix
	Total liabilities	FIN4		Audit opinion[1]	NFIN4	0 clean (unqualified), 1 otherwise
	Total equity	FIN5	CG metrics	Newcomer	NFIN6	1 yes, 0 otherwise
	Revenue	FIN6		Board Size	NFIN7	Count
	COGS	FIN7		CEO Duality[2]	NFIN8	0 no, 1 yes
	Gross profit	FIN8		Board Dependency[3]	NFIN9	0 independent,1 dependent
Financial ratios	Liquidity	FIN9		Board meetings	NFIN10	Count
	Profitability	FIN10		Audit committee size	NFIN11	Count
	Efficiency	FIN11		Ownership concentration	NFIN12	0 compliant, 1 non
	Leverage	FIN12		CEO tenure	NFIN13	0 compliant, 1 non
	Market ratio	FIN13		Adopted standards	NFIN14	0 SOCPA, 1 IFRS
			Auditor	Auditor status[4]	NFIN15	0 currently unauthorised, 1 authorised
				Audit Big 4[5]	NFIN16	0 no, 1 yes
				Auditor allegations[6]	NFIN17	0 no, 1 yes

1. There are four types of **auditor opinions**. The most common opinion is the Unqualified (Clean) Opinion, where the auditor believes the FSs are accurate and comply with accounting standards. This value is represented by 0. The second opinion is issued when the FSs are mostly accurate, but there is a specific issue that does not comply with standards, often explained in detail. This is called the Qualified Opinion. The remaining two types of opinions represent two levels of FS's misrepresentation [28]. To simplify the model's input, any opinion other than the clean opinion is represented as 1.
2. **CEO duality** indicates if the Chief Executive Officer holds the position of chairman of the board of directors. This feature is one of the main CG influencers, as holding both roles can lead to a concentration of power and potentially affect the board's ability to oversee management independently. This CEO duality takes 1 if the CEO is also the Chairman of the Board (indicating CEO duality exists) and 0 if the CEO and Chairman roles are held by separate individuals (no CEO duality).
3. According to the Saudi regulations by the CMA [13], CG regulations demand that at least one-third of the board members must be independent. Independent directors are non-executives. All independents are non-executives, but not all non-executives are independents (they might have other relationships). Therefore, the threshold in our design is 33%, so at least 33% must be independent. If less than 33%, then 1 (dependent). If 33% or more, then 0 (independent). **Board dependency** is measured using the following formula: *BD = (independent directors/ directors) *100*
4. Because the **status of the auditing firm** in the present indicates some flags about its integrity and professionalism, the performance history by allegations and by the status of the auditing firm by the authorities is highly effective in the analysis. If the audit firm (for a specific year) is authorised, this value takes 0, otherwise, the value is 1 if the audit firm is suspended in the current year.
5. The **size of the auditing firm** is considered as an important auditor's reputation feature as indicated by the findings of [2]. This variable asks whether the auditing firm is considered one of the Big 4 firms in Saudi Arabia (KPMG, PWC, EY, and Deloitte). It takes 1 if one of the big 4s, 0 otherwise.
6. It leverages the audit firm if it has no **previous allegations** or lawsuits either with companies or the authorities. The variable takes 1 if the auditing firm has no previous or current allegations, otherwise, it is 0.

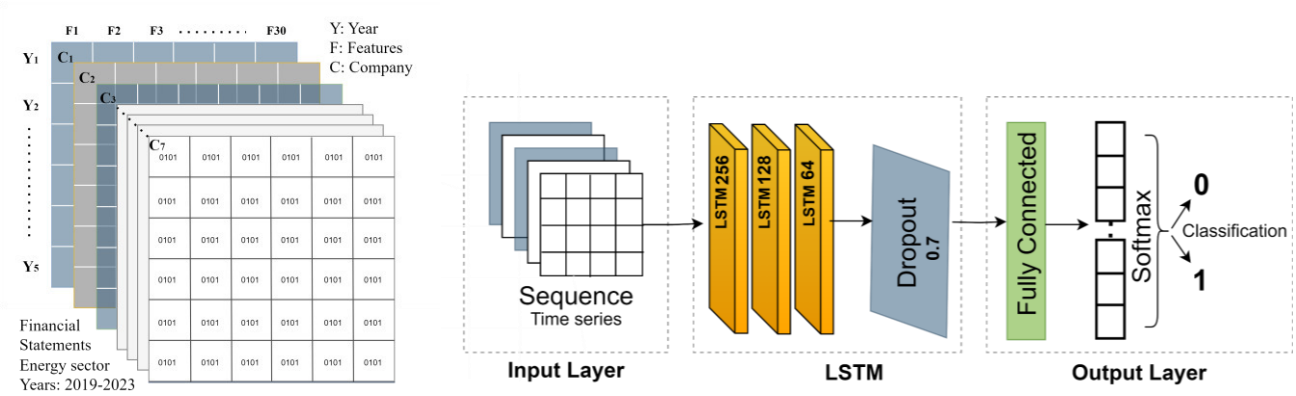

Fig. 3: Data Frame and Model Structure

model to capture temporal trends and patterns relevant to detecting CA. Time-series input can be mathematically expressed as follows:

$$X = \{x^1, x^2, x^3, \dots, x_T\}, x_t \in R^n \quad (1)$$

Where:
T number of time steps (years).
n number of features (financial + non-financial variables).
$\boldsymbol{x_t}$ feature vector at time step.

LSTM Layers

Multiple layers learn long-term dependencies between years of FSs, treating them as time series data. By leveraging memory cells and gated mechanisms, LSTM layers enable the model to identify changes in financial behaviour, such as trends in earnings manipulation or shifts in governance-related variables, that span across reporting periods. This is critical for CA detection, which may evolve gradually over time rather than appearing as sudden anomalies in a single year. The model consists of three LSTM layers with 256, 128, and 64 units, respectively. The first layer outputs the hidden states for all time steps to capture full temporal dependencies with 'sequence' output mode. The second layer with the 'last' output mode was used to summarise the input sequence into a single context vector, and the last layer with the same mode was used to further condense the sequential representation. The following equations describe the standard LSTM operation [29].

$$f_t = \sigma(Wfxt + Ufht - 1 + bf) \quad (2\text{-a})$$
$$i_t = \sigma(Wixt + Uiht - 1 + bi) \quad (2\text{-b})$$
$$o_t = \sigma(Woxt + Uoht - 1 + bo) \quad (2\text{-c})$$
$$\acute{c}_t = \tanh(Wcxt + Ucht - 1 + bc) \quad (2\text{-d})$$
$$c_t = f_t \odot c_{t-1} + i_t \odot \acute{c}_t \quad (2\text{-e})$$
$$h_t = o_t \odot tanh(c_t) \quad (2\text{-f})$$

Where:
σ: sigmoid activation $\boldsymbol{h_t}$: hidden state (output)
tanh: hyperbolic tangent activation $\boldsymbol{c_t}$: cell state
⊙: element-wise multiplication
W,U,b : trainable parameters

Dropout Layer

Since the model is trained on simulated data that may not fully reflect the variability of real-world financial statements, a dropout layer with a rate of 0.7 is applied to prevent data overfitting by randomly deactivating 70% of neurons during training. This regularises the LSTM model and reduces the risk of memorising synthetic patterns that do not generalise well to actual data.

Fully Connected Layer

This layer, with 2 neurons, aggregates the learned temporal features extracted by the preceding LSTM layers and maps them to the target output classes. By combining information across all time steps and input variables, the fully connected layer enables the model to produce a binary prediction that reflects the likelihood of CA behaviour in a company's financial statements.

$$z = W_{fc} h_T + b_{fc} \quad (3)$$

Where:
$\mathbf{W_{fc}}$: weight matrix for the dense layer
$\mathbf{b_{fc}}$: bias vector

SoftMax Layer

Performs binary classification (two classes, CA and N-CA). It consists of a single neuron with a sigmoid activation function, which outputs a probability score between 0 and 1, indicating the model's confidence that the financial statement exhibits CA behaviour. The threshold applied to convert this probability into a final class label is 0.5.

$$\bar{y} = \sigma(z) = \frac{1}{1 + e^{-z}} \quad (4)$$

Where:
$\bar{y} \in (\mathbf{0}, \mathbf{1})$ is the predicted probability of CA
A threshold $\boldsymbol{\theta = 0.5}$ is used to convert probabilities to class labels

$$\textbf{Predictions} = \begin{cases} 1 \; if \; \bar{y} > 0.5 & CA \\ 0 \; otherwise & N - CA \end{cases}$$

Training configuration

The model was trained using the Adam Optimiser for up to 100 epochs with a small mini-batch size of 10 to suit the limited number of training samples. Validation was performed every 5 epochs using a separate validation set, and early stopping was applied with a patience of 20 validation checks. Conversely, evaluation is performed once while training the model and another time during the testing phase, while comparing results. This phase is performed only once to assess the simulation efficiency.

F. Evaluation Metrics

An appropriate evaluation criterion must be established to assess the generalisation phase of both models and compare their performances. In the long-term plan of this research, the assessment criterion is to employ qualitative assessment, namely expert evaluation, to interpret testing results and validate the model's utility. This current stage relies on quantitative distributional analysis for the model outputs. Specifically, *SoftMax* confidence distributions are examined alongside the predicted labels to evaluate each model's behaviour and how biased it is towards each class. This approach allows for an initial assessment of how confidently the model classifies real-world FSs.

III. Implementation - The Experiment

This section details the implementation of CADM, focusing on the experimental process used to train and evaluate its performance. The implementation was carried out in the following phases: simulation (explained earlier), model training using the simulated datasets, model testing on real-world financial statements, and model evaluation as illustrated in Fig. 4 Training, Testing, and Evaluation phases are presented in the following subsections, outlining the experimental setup, hyperparameter choices, and evaluation outcomes.[22]

A. Training

As shown in Fig. 5, CADM1 achieved 100% accuracy in training and validation. In contrast, CADM2 achieved 92% accuracy in training and 80% accuracy in validation. All predicted labels were true by CADM1 and only 3 were false by CADM2 as shown in Fig. 6 These results were achieved after several iterations involving multiple rounds of dataset re-generation. The training datasets were carefully re-simulated to introduce variability and reduce the model's familiarity with recurring patterns, thereby promoting better generalisation and minimising overfitting.

To optimise model accuracy, various hyperparameters were fine-tuned, including the number of LSTM Units. They experimented with different neuron counts to determine the optimal balance between model complexity and learning capacity. While smaller configurations, such as 128, 64, and 32 units, were initially considered to reduce overfitting, the final architecture adopted a larger structure —256, 128, and 64 units —demonstrating improved learning performance without compromising generalisation. This architecture, applied to both CADM1 and CADM2, retained sufficient capacity to capture complex temporal dependencies in the data while maintaining robust performance on unseen samples, despite the relatively small dataset size.

Moreover, the Learning Rate was adjusted to prevent convergence issues, and the batch size and epochs were optimised for computational efficiency and model performance. Additionally, L2 regularisation was applied and a dropout layer to mitigate overfitting. Finally, the Optimiser Selection compared Adam, RMSprop, and SGD to identify the most effective optimisation strategy.

B. Testing

Following the training of both models, CADM1 and CADM2, the next step is to test them to assess their generalisation capabilities. Testing is designed to be implemented on

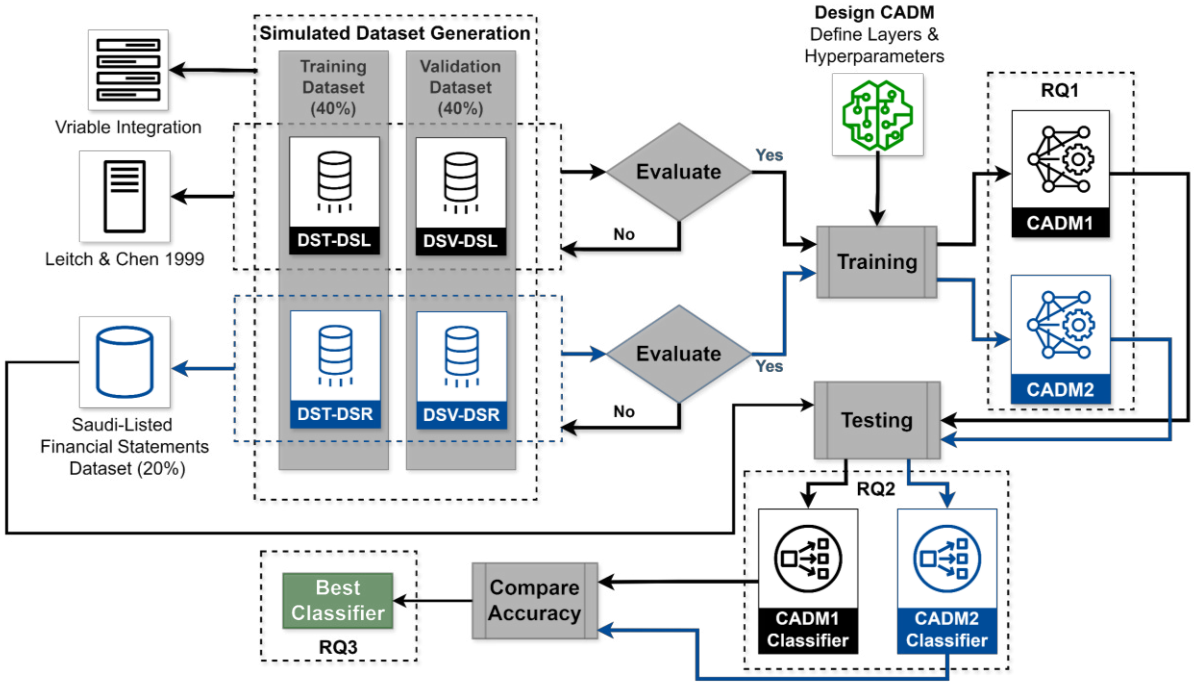

Fig. 4: Workflow of CADM Simulation, Training, Testing, and Evaluation

the real-world dataset; the collected and pre-prepared FSs, which have become DSR after preparation. Due to having two models, testing was done twice on the same dataset. However, DSR is prepared to be in the same shape as the simulated dataset to have correct and accurate testing and be ready to be used by CADM.

However, CADM1 predicted five out of six companies as class 1 (CA) Fig. 7, but with SoftMax probabilities clustered very close to the classification threshold (ranging between 0.5001 and 0.5042) Fig. 8. The single prediction of class 0 (N-CA) was also marginal, with a class 0 probability of just 0.5004. This narrow range of values indicates that CADM1 was uncertain in its predictions, which reflects the model's weak classification capability and lack of confidence when applied to real-world financial statements.

In contrast, CADM2 predicted all companies class 0 (N-CA) Fig. 7, with SoftMax probabilities ranging from 0.5021 to 0.5063. Being close to the decision threshold as well does not indicate any better performance than CADM1, although they have more consistency (all above 0.5). Still, CADM2 demonstrates modest and more stable confidence in labelling the data.

IV. Discussion

This study successfully developed and evaluated two LSTM-based models (CADM1 and CADM2) to classify FSs

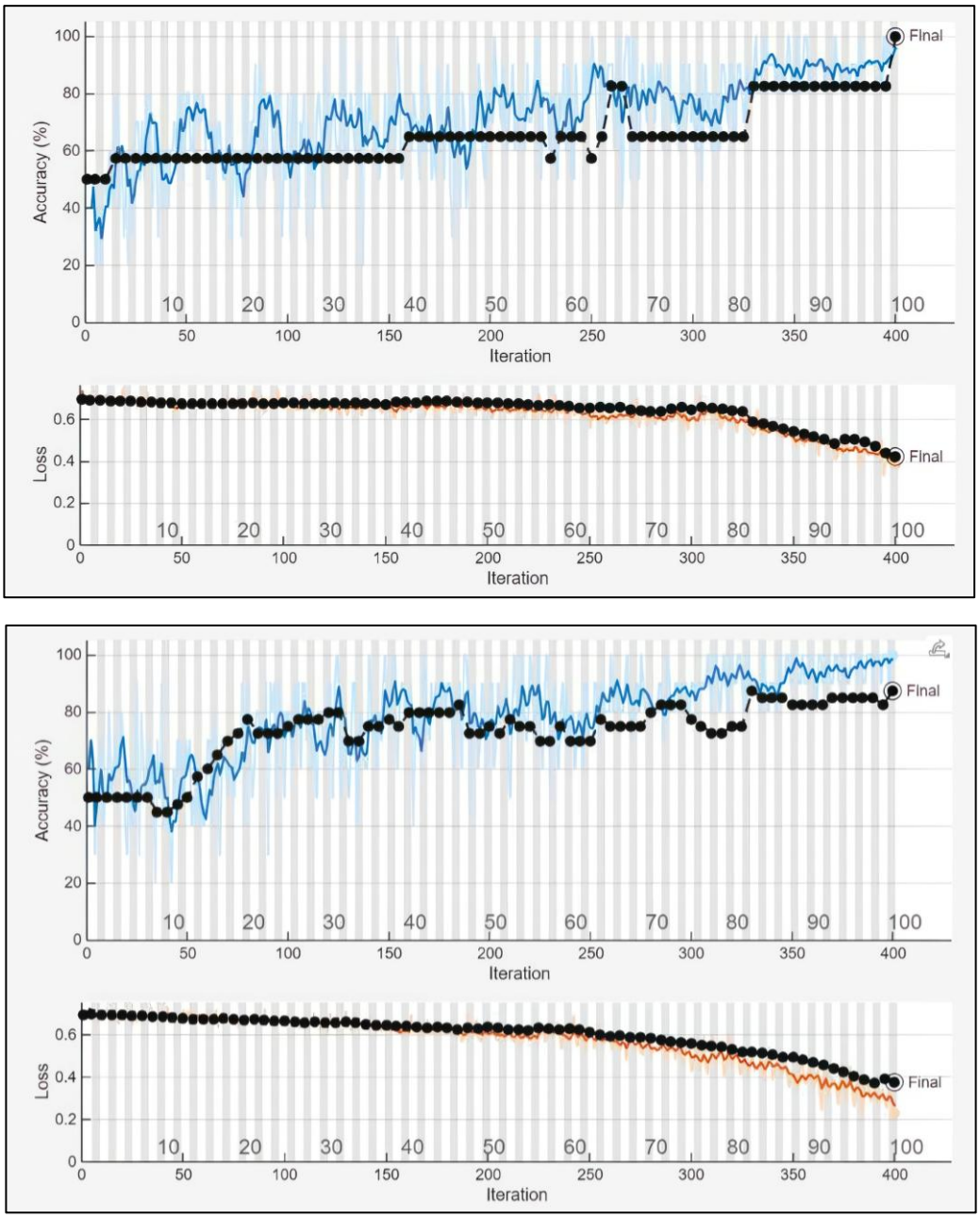

Fig. 5: Training accuracy and loss curves for CADM1 (top) and CADM2 (bottom), showing the models' learning performance over 100 epochs

based on the likelihood of CA practices being used in their preparation. The models were trained and validated to assess their effectiveness in controlled simulations and real-world Saudi market data. Additionally, the study compared the models' performance regarding the underlying simulation approach used to generate their training dataset. This section discusses the research objectives and demonstrates meaningful contributions to DL-driven financial analysis.

RQ1: Training CADM: Learning from Simulated Data

The results of training CADM reflect a key contrast between real-world context and theoretical modelling. CADM1 achieved perfect accuracy (100%) in both training and validation phases. This technical success confirms that the model architecture is responsive to clean data with well-defined patterns. However, this result may also indicate overfitting to the ideal dataset that lacks real-world variability.

In contrast, training CADM2 was more challenging. Despite extensive hyperparameter tuning, CADM2's accuracy was lower than CADM1's, reflecting the inherited complexity of real-world Saudi financial behaviour. This result is not a flaw but a strength: it underscores CADM2's exposure to more realistic noise, non-linear interactions, and imperfect relationships characterising actual FSs.

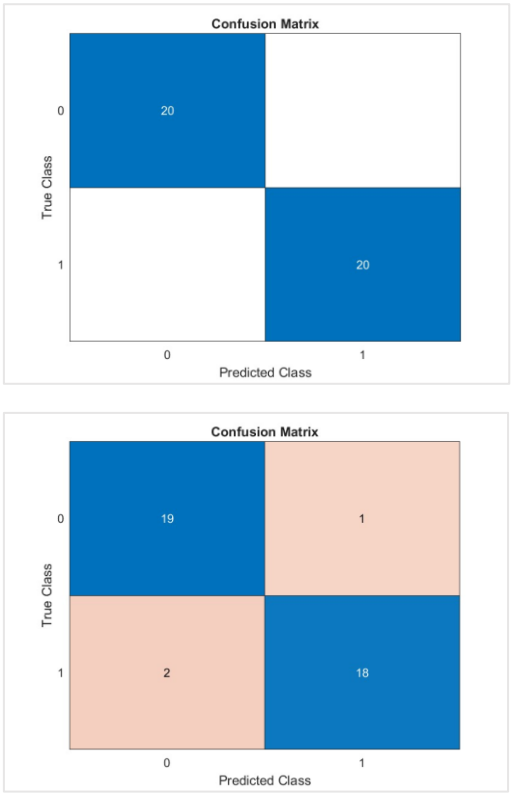

Fig. 6: Classification Results of CADM1 Training (top) and CADM2 Training (Bottom)

In addition, the training dynamics promote valuable reflections on the boundaries of CADM's optimal performance, raising questions about model robustness and the interpretability of results when using overly cleaned data, a significant debate in DL-based financial models. CADM1's perfect performance might not perfectly generalise to real-world data, while CADM2, despite lower learning scores, likely learned patterns closer to reality. This observation supports the argument in the literature about the risk of training financial AI models on idealised datasets.

RQ2: Generalisation: Performance on Real-World Data

As ground truth for CA is not available, innovative assessment methods become necessary. At this stage of the research, the generalisation capacity of both models was evaluated using class distribution patterns and feature-based differentiation. Although theoretically accurate, CADM1 was more liable to misclassification when confronted with real-world variance. In contrast, the final version of CADM2 testing showed that all companies in the Energy sector are CA-free. This can be interpreted as the model trained on a dataset that already sees these scenarios as non-CA, which reflects stronger generalisation capacity from CADM2.

From the perspective of CA theory, CA is intentionally designed to evade detection mechanisms. Consequently, a deep contextual understanding and the incorporation of non-financial signals are essential for effective detection. This may explain the comparatively better classification performance by CADM2 when tested on real-world FSs, as it appears to have

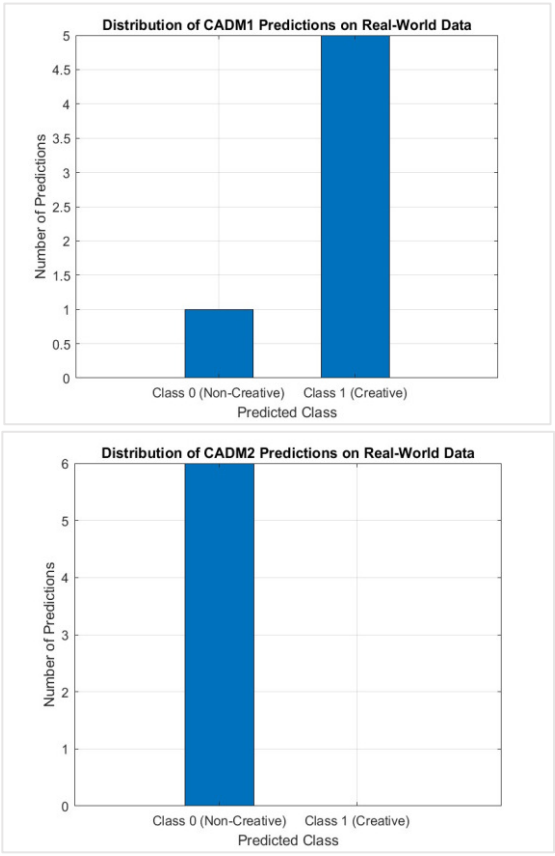

Fig. 7: Classification results of CADM1 (top) and CADM2 (bottom) on the real-world dataset, showing predicted counts for Class 0 (N-CA) and Class 1 (CA).

captured some attributes of the complexity inherent in real-world data. Nevertheless, generalisation was naturally challenged by sector-specific anomalies. Companies in this sector have diverse accounting scenarios and include a dominant outlier (ARAMCO), with significantly different scales and metrics compared to its peers. These factors influenced the model's sensitivity and underscore the importance of sector-aware tuning in future iterations.

RQ3: Evaluating The Simulation Base: Theory vs Reality

The comparison between CADM1 and CADM2 provides critical insight into how simulation strategy affects model behaviour. Evaluating the simulation base was done by comparing the results of CADM1 and CADM2 classifiers. The SoftMax probability distributions further illustrate these tendencies. While CADM1 demonstrated theoretical strength, CADM2 has more consistent confidence values, particularly when they were consistently above 0.5 and demonstrated more stable confidence in labelling companies as N-CA. The more consistent confidence profile of CADM2 is attributed to its training on a dataset simulated using real-world statistical analysis, which enabled it to better mirror the characteristics of actual financial reporting in the Saudi market. This confirms the validity of using real-world statistical patterns in training simulations.

The contrasted results between CADM1 and CADM2 illustrate that models trained on idealised literature-driven simulations achieve higher accuracy yet lower generalisation ability to real-world FSs compared to models trained on real-world-based simulations. This supports the theoretical argument in the literature that training models on clean data may not adequately prepare them for the complexities and imperfections found in real-world data [30], [31]. In addition, the experiment in this study suggests that hybrid approaches, such as blending literature-based metrics with real-world features, may offer the most promising path forward.

V. Conclusion

This study represents a significant milestone in developing a data-driven framework for detecting CA within Saudi-listed companies. By training two variants of the Creative Accounting Detection Model (CADM) using different simulation criteria, this research establishes a credible foundation for AI-enabled financial monitoring. The model was then tested on the real-world dataset collected from the Saudi market portal to check if the model is capable of looking at how financial and non-financial metrics evolve, CADM and classifying Saudi-listed companies in the Energy sector as creative accounting or non-creative accounting practitioners.

Both SoftMax confidence distributions of CADM predictions on the real-world dataset models offer a significant insight: the importance of strong theory and the need to reflect real-world scenarios when detecting financial anomalies. The implication of this study extends beyond model performance. It sets a new direction for applying deep learning to complicated grey areas like creative accounting. Unlike fraud detection, this model identifies forms of financial misrepresentation that fall within regulatory bounds but still distort finan-

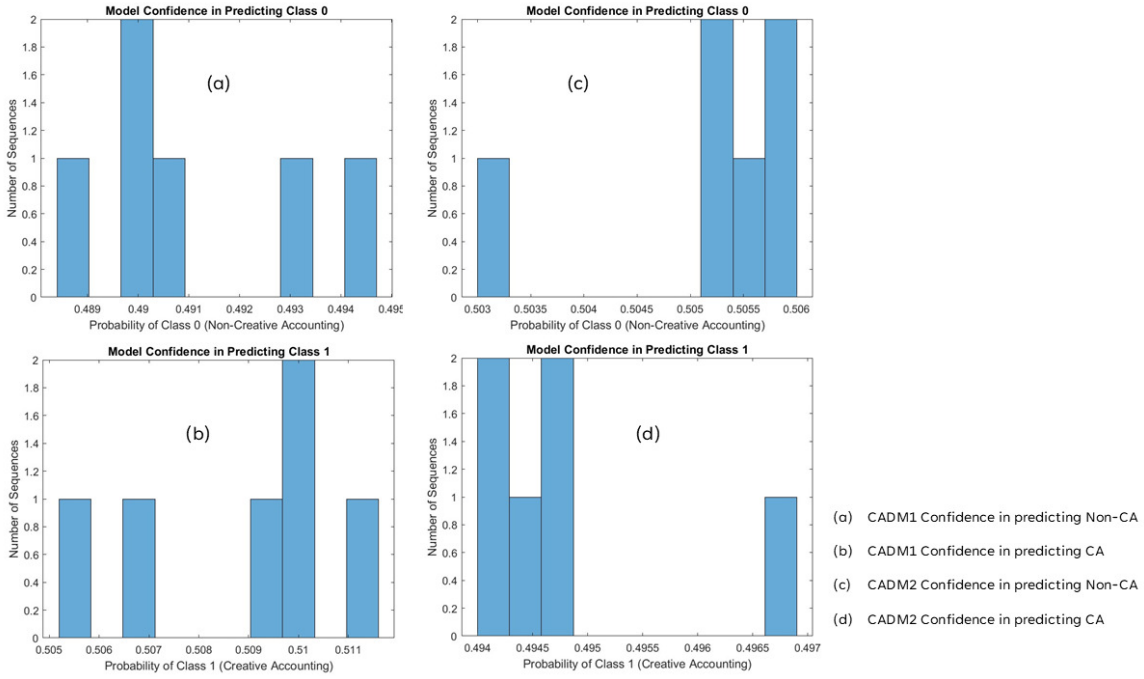

Fig. 8: SoftMax confidence distributions of CADM predictions on the real-world dataset

cial reality. Further research may involve integrating additional financial indicators such as the revenue recognition, misclassification of accounting elements, and off-balance sheet financing, and governance indicators, such as management compensations and narrative disclosure. It could also add explainability mechanisms to the model, such as applying custom attention layer to the model to enhance model transparency and make CADM a valuable tool for regulators and auditors. In addition, the simulation scope could be expanded to have semi-creative accounting data, and the testing could include other sectors to improve sensitivity and real-world relevance. Future work is intended to integrate expert interpretation of classification results through interviews to validate the model's output and illuminate practical constraints.

This research serves as a foundation for the next generation of intelligent accounting analytics. Besides considered a proof of the feasibility of AI in financial transparency, it promotes meaningful integration of simulation, sector knowledge, and regulatory insight.

References

[1] L. Revsine, 'Enron: Sad but inevitable', *J. Account. Public Policy*, vol. 21, no. 2, pp. 137–145, 2002, doi: 10.1016/S0278-4254(02)00044-3.

[2] Hoje Jo, Annie Hsu, Rosamaria Llanos-Popolizio, and Jorge Vergara-Vega, 'Corporate Governance and Financial Fraud of Wirecard', *Eur. J. Bus. Manag. Res.*, Mar. 2021, doi: 10.24018/ejbmr.2021.6.2.708.

[3] O. Mesioye and I. Bakare, 'Evaluating Financial Reporting Quality: Metrics, Challenges, and Impact on Decision-Making', *Int. J. Res. Publ. Rev.*, vol. 5, pp. 1144–1156, Oct. 2024, doi: 10.55248/gengpi.5.1024.2735.

[4] M. Bineid and A. Asiri, 'Creative Accounting Incentives and Techniques in Saudi Public Companies', *KAU J. Adm. Econ.*, vol. 2, no. 27, 2013, doi: 10.4197/Eco.

[5] Jones Michael, *Creative accounting, fraud, and international accounting scandals*. Chichester, West Sussex, England; John Wiley and Sons, 2011.

[6] E. E. Akpanuko and N. J. Umoren, 'The influence of creative accounting on the credibility of accounting reports', *J. Financ. Report. Account.*, vol. 16, no. 2, pp. 292–310, 2018, doi: 10.1108/JFRA-08-2016-0064.

[7] M. Bineid, A. Khanina, N. Beloff, and M. White, *Integrating Non-financial Data into a Creative Accounting Detection Model: A Study in the Saudi Arabian Context*, vol. 504. in Lecture Notes in Business Information Processing, vol. 504. Cham: Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-61657-0.

[8] M. Bineid, N. Beloff, M. White, and A. Khanina, 'CADM: Big Data to Limit Creative Accounting in Saudi-Listed Companies', presented at the 18th Conference on Computer Science and Intelligence Systems, Sep. 2023, pp. 103–110. doi: 10.15439/2023F3888.

[9] M. D. Beneish, 'The Detection of Earnings Manipulation', *Financ. Anal. J.*, vol. 55, no. 5, pp. 24–36, 1999, doi: 10.2469/faj.v55.n5.2296.

[10] M. D. Beneish and P. Vorst, 'The Cost of Fraud Prediction Errors', *Account. Rev.*, vol. 97, no. 6, pp. 91–121, Oct. 2022, doi: 10.2308/TAR-2020-0068.

[11] P. M. Dechow, A. P. Hutton, J. H. Kim, and R. G. Sloan, 'Detecting Earnings Management: A New Approach', *J. Account. Res.*, vol. 50, no. 2, pp. 275–334, 2012, doi: 10.1111/j.1475-679X.2012.00449.x.

[12] J. Wang, S. Hong, Y. Dong, Z. Li, and J. Hu, 'Predicting stock market trends using LSTM networks: overcoming RNN limitations for improved financial forecasting', *J. Comput. Sci. Softw. Appl.*, vol. 4, no. 3, pp. 1–7, 2024.

[13] A. Abbasi, C. Albrecht, A. Vance, and J. Hansen, 'METAFRAUD: A META-LEARNING FRAMEWORK FOR DETECTING FINANCIAL FRAUD', *MIS Q.*, vol. 36, no. 4, pp. 1293–1327, 2012, [Online]. Available: http://www.misq.org

[14] P. Craja, A. Kim, and S. Lessmann, 'Deep learning for detecting financial statement fraud', *Decis. Support Syst.*, vol. 139, Dec. 2020, doi: 10.1016/j.dss.2020.113421.

[15] M. Schreyer, T. Sattarov, C. Schulze, B. Reimer, and D. Borth, 'Detection of Accounting Anomalies in the Latent Space using Adversarial Autoencoder Neural Networks', Aug. 02, 2019, *arXiv*: arXiv:1908.00734. doi: 10.48550/arXiv.1908.00734.

[16] F. Alam and S. Ahmad, 'Intelligent Fraud Detection Framework for PFMS Using HGRO Feature Selection and OC-LSTM Fraud Detection Technique', *SN Comput. Sci.*, vol. 4, no. 4, p. 400, May 2023, doi: 10.1007/s42979-023-01855-5.

[17] Yanash Azwin Mohmad, 'Credit Card Fraud Detection Using LSTM Algorithm', vol. 1, no. 3, 2022, doi: https://doi.org/10.31185/wjcm.60.

[18] *Deep Learning-Based Corporate Performance Prediction Model Considering Technical Capability*. [Online]. Available: https://www.mdpi.com/2071-1050/9/6/899

[19] H. A. Almustawfiy, 'Creative Accounting Applications, Oppotunistic Behavior, and Integrity of Accounting Information System: The Case of Iraq', vol. 24, no. 6, pp. 1–12, 2021.

[20] M. Al Shetwi, 'Earnings Management in Saudi Nonfinancial Listed Companies', *Int. J. Bus. Soc. Sci.*, vol. 11, no. 1, 2020, doi: 10.30845/ijbss.v11n1p3.

[21] A. F. Al-Hassan, 'Earnings Management using Accruals: Empirical Study on Saudi Companies', *Arab. J. Adm.*, 2018, [Online]. Available: http://search.mandumah.com/record/940867

[22] R. Mubeen, D. Han, J. Abbas, S. Álvarez-Otero, and M. S. Sial, 'The Relationship Between CEO Duality and Business Firms' Performance: The Moderating Role of Firm Size and Corporate Social Responsibility', *Front. Psychol.*, vol. 12, p. 669715, Dec. 2021, doi: 10.3389/fpsyg.2021.669715.

[23] Saul Calderon-Ramirez, Shengxiang Yang, David Elizondo, 'Semi-supervised Deep Learning for Image Classification with Distribution Mismatch: A Survey', *IEEE Trans. Artif. Intell.*, 2022.

[24] R. A. Leitch and Y. Chen, 'Simulation of Controlled Financial Statements', *Rev. Quant. Finance Account.*, vol. 13, no. 2, pp. 189–207, Sep. 1999, doi: 10.1023/A:1008304127780.

[25] I. D. L. TORRE, *Creative Accounting Exposed*. Palgrave Macmillan London, 2008.

[26] C. W. Mulford and E. E. Comiskey, *The Financial Numbers Game: Detecting Creative Accounting Practices*. John Wiley & Sons; 1st Edition, 2002.

[27] F. Tassadaq and Q. A. Malik, 'Creative accounting and financial reporting: Model development and empirical testing', *Int. J. Econ. Financ. Issues*, vol. 5, no. 2, pp. 544–551, 2015.

[28] CMA, 'The Most Prominent Observations On The Disclosures Of The Financial Statements Of Listed Companies', 2021 2020.

[29] S. Hochreiter and J. Schmidhuber, 'Long Short-Term Memory', *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[30] V. Kamath, R. A., V. G. Kini, and S. Prabhu, 'Exploratory Data Preparation and Model Training Process for Raspberry Pi-Based Object Detection Model Deployments', *IEEE Access*, vol. 12, pp. 45423–45441, 2024, doi: 10.1109/ACCESS.2024.3381798.

[31] Y. Bengio, Y. Lecun, and G. Hinton, 'Deep learning for AI', *Commun. ACM*, vol. 64, no. 7, pp. 58–65, Jul. 2021, doi: 10.1145/3448250.