Truth Under Pressure: A Deep Learning-Based Lie Detection System for Online Lending Using Voice Stress and Response Latency

Ahmad Ihsan Farhani[1], Alhadi Bustamam[2*], Rinaldi Anwar[3], Dra. Titin Siswantining[4]
Department of Mathematics-Faculty of Mathematics and Natural Sciences, Universitas Indonesia, Depok, Indonesia [1, 2, 4]
Data Science Center (DSC)-Faculty of Mathematics and Natural Sciences, Universitas Indonesia, Depok, Indonesia [1, 2]
Global Risk Management (GRM), Jakarta, Indonesia [3]

***Abstract*—The rapid increase in defaults in the online lending industry highlights significant flaws in current debtor verification, which largely relies on static, preparable interviews, leading to high non-performing loans. Existing research is fragmented: while Large Language Models (LLMs) show promise in question generation, their application is confined to non-financial domains like education, and lie detection studies often analyze modalities in isolation. This study addresses this critical gap by proposing the first *integrated* AI-driven system for this context. We solve the problem in two parts: 1) A Llama 3 LLM is fine-tuned to generate dynamic, biodata-tailored questions, preventing the rehearsed answers that plague static interviews. 2) A novel multimodal deep learning model is developed to analyze the response, uniquely fusing vocal acoustic features and response latency—two key deception indicators that prior work has failed to combine. The Llama 3 model produced a low perplexity score (2-3), and the lie detection model achieved 70% testing accuracy with a 70.9% F1-Score. Despite signs of overfitting, this framework provides a novel, intelligent decision-support tool to reduce fraud and manage default risks more effectively.**

***Keywords*—Online lending; lie detection; large language model; deep learning; voice acoustics; response latency**

I. Introduction

In 2024, Indonesia's internet penetration was around 79.50%, which is about 221.5 million people [1]. This high percentage indicates a high level of connectivity and implies a propelling growth in various digital innovations, especially in financial technology (fintech) services. Peer-to-peer (P2P) lending, commonly known as online lending, is one of the most dynamic segments within fintech. This service allows lenders and debtors to transact directly through digital interfaces. Online lending users grew tremendously from 2.7 million to 8.8 million within a year in 2024 [2]. Post-COVID-19 economic realities and a heightened digital consumption mindset have further spurred the adoption of these services [3].

However, the availability of online loans has led to increasing irresponsible borrowing behavior. An increasing number of debtors tended to skip out on easy-to-access loans, which in turn resulted in a rise in the "non-performing loans" (NPLs). In the context of P2P lending, this is measured by the 90-Day Default Rate (TWP90), which refers to loan repayment that has defaulted beyond 90 days [4]. Based on the latest information from the Financial Services Authority of Indonesia, stagnant online loans jumped significantly from IDR 785.94 billion in January 2022 to IDR 1.9 trillion by June 2024 [5]. This is illustrated in Fig. 1.

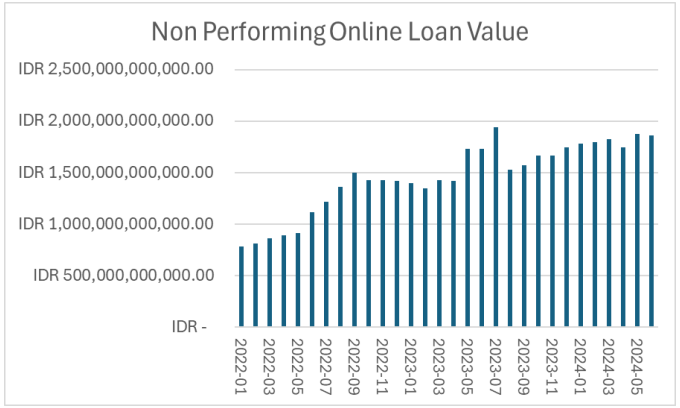

Fig. 1. Accumulated defaulted online loans, 2022–2024.

Moreover, as shown in Fig. 2, the 19 to 34 age group is the largest contributor to loan defaults, followed by individuals aged 35 to 54 [6]. This trend raises serious concerns because the 19 to 34 demographic is the economically productive part of the population that should ideally drive economic growth.

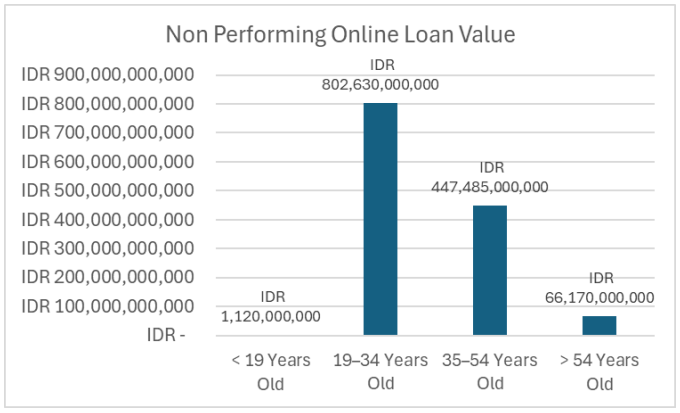

Fig. 2. Default rate by age group.

One of the key factors causing the high rate of loan defaults among the younger productive generations is their

*Corresponding author.

low financial literacy. Several young people use online lending platforms without truly understanding the risks [7]. In 2024–2025, the Financial Services Authority conducted a survey called Indonesia's National Financial Literacy Survey to measure the financial literacy of the Indonesian population. The result of the survey is that in Indonesia, the financial literacy index is only around 65%–66%, and for digital financial literacy, it is at 62% [8],[9].

The survey results from [7]-[9] imply that young people in Indonesia might lack financial understanding, which increases the risk of falling victim to online lending schemes. Several debtors have limited understanding and do not fully understand how online lending works. This causes the debtors to not fully understand the risks and dangers of illegal online lending practices [10]. If this situation continues, the number of non-performing loans will continue to increase. The increase in non-performing loans will weaken personal finances, decrease household consumption, decrease investment, and limit entrepreneurship growth. In the long term, this situation will threaten the national economy.

In addition, verification procedures in online methods of obtaining loans mainly consist of filling in personal data without conducting a direct interview. This opens the opportunity for dishonest debtors to exploit this system. Dishonest debtors can manipulate or change the biodata details they submit, increasing the chances of online loan default. Even in cases where interviews are held, asking static, pre-set questions can be inadequate, as it allows the debtors to prepare rehearsed answers that may not truly represent their financial situation.

To reduce this problem, it is recommended that interview questions be designed to reflect the details of the debtor's biodata so that questions are specific to the debtor and have less predictable answers. However, manual operation on this personalized level would be labor-intensive. In this study, we propose a large language model (LLM)-based automatic interview system and integrate it with a multimodal deep learning architecture.

To provide on-the-spot responses without having to prepare or memorize answers, LLM will use automated question generation (AQG) to generate personalized questions tailored to the debtor biodata. A multimodal model that assesses vocal acoustic characteristics and response latency as possible markers of dishonesty was used to analyze the responses. This system is intended to be implemented before a loan is approved to improve verification accuracy and lower the risk of loan default.

Even though it is still used in the educational domain, previous research has shown the potential of LLMs such as T5, mT5, and Llama in generating questions [11]-[13]. Based on previous research, one of the signs of lying is a change in the acoustic characteristics of speech, such as an increase in pitch and speech disfluency; features such as longer answering process are also one of the deception clues [14]-[16]. All of these signs of deception must be combined to improve model performance as different modalities require a multimodal mechanism. This multimodal model will classify the answer of the debtor as truth or lie.

The major contributions of this study are as follows:

- Proposes a novel integration of vocal acoustic features and response latency to enhance deception detection. These two elements are typically analyzed separately.
- Utilizes a LLM to perform AQG within the financial domain, effectively simulating the online lending interview process.
- Introduces a new verification approach aimed at reducing NPL rates by combining LLM driven questioning and deception detection techniques for a prospective debtor.
- Develops a cost-effective and scalable deception detection framework that supports improved decision-making and operational efficiency in online lending platforms.

This study is organized into five sections. Section I provides the study's background and motivation. Section II explains and shows some related works and previous research relevant to this topic. Section III explains the dataset, methodology, and models used in this study. Section IV presents and discusses the results. Finally, Section V concludes the study and outlines potential directions for future work.

II. Literature Review

This section outlines previous studies related to the current study. This section is divided into two subsections: AQG using LLM and lie detection.

A. AQG Using LLM

AQG using LLM has seen significant growth, particularly in the education domain. As shown in the study [11] and [12], the LLM model with mT5 and T5 architecture has the ability to generate questions that are appropriate to the context of the given input with good human language and good readability. In [11], the authors used the mT5 model trained on the TQuADv2 and XQuAD datasets to generate comprehension questions, achieving BLEU-1 of 47.6 and ROUGE-L of 53.9. Similarly, EduQG, a QG model developed by [12] using T5-small and T5-base, reached BLEU-4 of 15.94 and F1 score of 33.12.

More recently, [13] fine-tuned LLaMA-2 (7B parameters) for question–answer extraction and generation on the SQuAD and PBE datasets, achieving a BLEU-4 of 23.53 and a METEOR of 66.3. Previous studies have shown that LLM can generate dynamic, context-appropriate questions. However, its application is still limited to the educational domain and has not been further explored for adaptation in risk assessments, such as debtor eligibility interviews for online loans.

B. Lie Detection

Several studies have explored lie detection using various modalities such as text, audio, video, facial expression, and multimodal approaches. For instance, [17] implemented a LLM (FLAN-T5) for verbal lie detection using datasets such as Deceptive Opinions, Hippocorpus, and the Intention Dataset, achieving an accuracy of 82.72%. The architecture

created by [18] was named DeepLie, where facial features were extracted using face detection and classified using a deep learning model. The developed DeepLie model achieved a performance of 81.82% on a real-world trial dataset.

Different research approach by [19] extracted high-level features from videos using tools such as OpenFace, part-of-speech n-grams, MFCC, and Covarep and reported a test AUC of 0.730. In [20], the authors combined video, audio (MFCC), and text (GloVe) features and used machine learning classifiers, such as SVM, K-SVM, Decision Tree, and Random Forest, achieving 87.73% accuracy on a trial dataset. Focusing on audio-based detection, [21] developed an acoustic lie detector using MFCC features and models such as SVM and SGD, achieving 62.4% accuracy on the Columbia-SRI-Colorado dataset. In [22], the authors proposed an explainable LSTM-based model that leverages stress analysis in voice signals, achieving an accuracy of 92.4% on a stress voice dataset.

Multimodal deep learning also shows promising results. Research conducted by [23] utilized range of deep learning architecture such as 3D CNNs, text CNNs, and multilayer perceptron and combined them with OpenSMILE for audio features. The research combines video, transcript, and speech to classify a video into a lie or truth class, achieving an impressive accuracy of 97.99% on a trial dataset. Similarly, [24] applied OpenFace and FACS with LSTM for deception detection in videos across datasets such as Silesian, trial, and Bag of Lies, achieving an accuracy of 90.90%.

Another study conducted by [25] used cognitive signals, namely, response latency and error rates, when participants answered unexpected questions as material in detecting lies. Logistic regression, KNN, and SVM were the classifiers used in the study. The highest performance based on the study was an accuracy of 81%, with an AUC of 92%, recall of 80%, precision of 84%, and an F1 score of 80%, highlighting the potential of behavioral timing analysis in lie detection.

C. Research Gap and Proposed Contributions

The literature review reveals significant progress in both AQG and lie detection; however, it also highlights key limitations that this study aims to address.

First, in the field of Automated Question Generation (AQG), the application of powerful LLMs like T5, mT5, and LLaMA is in the educational domain [11]-[13]. There is a distinct lack of research exploring their use in dynamic, high-stakes financial risk assessments, such as debtor interviews.

Second, in Lie Detection, some studies focus solely on a single modality, such as text [17], facial expressions [18], [24], or audio features [21], [22]. Some multimodal systems often combine mostly video, text, and audio either all three of modalities nor combination of 2 modalities [19], [20], [23]. The specific integration of vocal acoustic features and response latency is not explored yet, even though both of the modalities has strong indicators of the cognitive effort associated with deception. Mostly both feature are analyzed separately.

This study directly addresses these gaps. We propose a system that: 1) adapts an LLM for AQG within the financial (online lending) domain and 2) develops a novel, streamlined multimodal model that specifically combine both vocal acoustics and response latency. The following TABLE Isummarizes this research gap and our proposed contributions.

Unlike previous AQG studies that primarily focused on educational or comprehension contexts, our approach introduces domain-specific adaptation and fine-tuning of a large language model for financial risk assessment. The novelty lies in three aspects:

- Prompt engineering strategy, each input prompt integrates structured debtor biodata with contextual financial indicators, enabling the model to generate dynamic and risk-relevant interview questions rather than general comprehension items.
- Domain adaptation, Llama 3 was fine-tuned on a synthetic dataset specifically designed to reflect debtor profiles, borrowing intents, and loan conditions, thus aligning the model's linguistic generation with financial verification contexts.
- Automated reasoning constraint, during decoding, we applied controlled generation parameters (low temperature and limited Top-K sampling) to ensure factual consistency and reduce speculative or irrelevant question output.

Together, these innovations allow the model to perform adaptive, context-aware question generation that simulates human financial interviews a task previously not explored in prior AQG research.

TABLE I. SUMMARY OF RESEARCH GAPS AND PROPOSED CONTRIBUTIONS

Research Area	Identified Limitations / Gaps in Previous Studies	Proposed Solution in This Study
AQG	Application is heavily restricted to the educational domain [11]-[13]. Lacks adaptation for financial risk assessment or debtor interviews.	Develop and fine-tune a LLM model to generate dynamic, personalized interview questions based on debtor biodata specifically for the online lending context.
Lie Detection	Some studies focus on a single modality, such as text [17], facial expressions [18], [24], or audio features [21], [22]. Existing multimodal systems combine video, text, and audio [19], [20], [23]. The specific integration of vocal acoustic features and response latency is not explored. These features are typically analyzed separately.	Propose a novel and efficient multimodal deep learning model that integrates two inputs: vocal acoustic features and response latency.

III. METHODOLOGY

The research methodology follows a multi-stage workflow, as illustrated in Fig. 3, progressing from data acquisition to model integration. The process begins with two

sequential stages: Data Collection and Data Preprocessing. In the collection stage, two distinct datasets are gathered: a synthetic dataset of borrower biodata and interview questions for the AQG task, and the courtroom trial video dataset [26] for lie detection. The subsequent preprocessing stage then cleans, formats, and transforms both the text and audio datasets to make them suitable for training. Following these preparatory steps, the process bifurcates into two distinct, parallel development tracks.

The first track is dedicated to the Automated Question Generation (AQG) component, which involves fine-tuning the Llama 3 model on the prepared biodata dataset. Simultaneously, the second track focuses on building the lie detection system, which entails training the multimodal deep learning model on extracted vocal characteristics and response timing data. Once training is complete, each model is independently assessed in the Model Evaluation stage: the LLM is measured using perplexity and expert review, while the lie detection model is validated using classification accuracy and F1 score. Finally, the two validated models are brought together in the Implementation stage, where they are integrated into a single, cohesive framework for deployment (see Fig. 3).

A. Data Collection

This subsection discusses the data collection process and the datasets used in the study. The dataset used in the research consists of two different data: 1) data for training the LLM model to generate interview questions and 2) data used to train the lie detection model. Each dataset serves a distinct purpose in supporting the development of a comprehensive and intelligent interview system for online lending verification.

1) Interview question dataset: To the author's knowledge, there is no public data containing the debtor's biodata and interview questions that correspond to them, not even the debtor's personal data. This is because such data are sensitive and will not be publicly available. Therefore, this study uses the synthetic data generated using GPT-4o and GPT-4o mini.

The resulting dataset includes 595 borrower profiles and 2,713 generated questions. The data is organized into three key components: Biodata, which contains the prospective debtor's personal and financial information; Analysis, which provides contextual interpretation of the biodata; and Questions, which comprises tailored interview questions generated based on the analysis. A sample of the dataset is shown in 0. This structured format ensures that both question generation tasks are supported by the dataset.

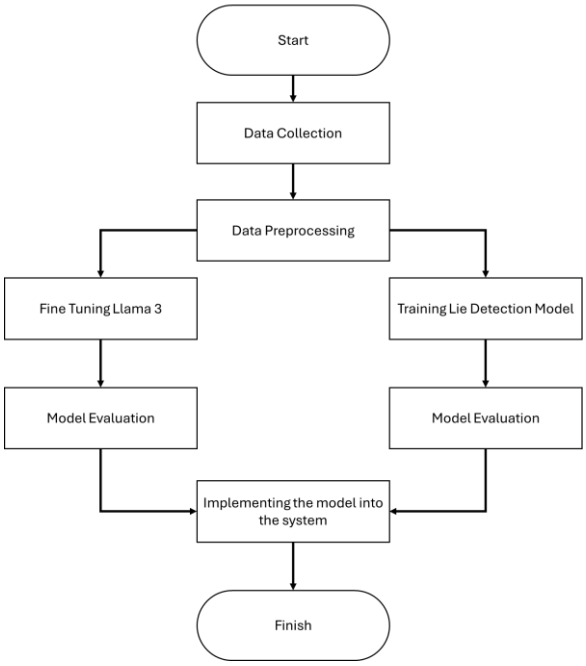

Fig. 3. Methodology of the study. After preprocessing, the fine-tuning of Llama 3 (left) and the training of the lie detection model (right) run in parallel.

TABLE II. SAMPLE BIODATA OF THE QUESTION GENERATION DATASET

Biodata	Analysis	Question
Full Name: Rudi Hartono; **Province**: West Java; **City**: Bekasi; **Educational Background**: High School; **Marital Status**: Single; **Reason for Loan**: Household needs; **Residence**: Rental; **Private Vehicle**: Motorcycle; **Company Name**: PT. Sentosa Abadi; **Occupation**: Private employee Industry; **Sector**: Industry/Factory; **Position**: Supervisor; **Net Monthly Income**: Rp 8,500,000; **Household Monthly Income**: Rp 8,500,000; **Years of Employment**: 4 years; **Emergency Contact**: Parent; **Emergency Contact Name**: Sri Hartono; **Loan Amount**: Rp 15,000,000; **Loan Tenure (Months)**: 18 months; **Monthly Interest Rate**: 6.0%	The reason for borrowing for household needs indicates a shortfall in meeting basic necessities, which may suggest potential financial constraints or inadequate budget management.	1. Do urgent household needs often consume a significant portion of your income? 2. Why do you feel the need to apply for a loan to cover your basic household needs? 3. Do you experience difficulties in managing your monthly finances for daily needs?

2) Lie detection dataset: This study employs the courtroom trial video dataset by [26] for lie detection. The dataset consists of 121 video clips (61 videos are labeled as deceptive class and 60 videos are labeled as truthful class). The videos are labeled based on court evidence and rulings. Audio and these models, which include millions or even billions of parameters, are trained on vast amounts of text data and have demonstrated remarkable capabilities in understanding and generating human-like text. Response latencies were extracted from the videos for analysis. The lie detection dataset compiled from 56 individuals consists of 21 females and 35 males with ages ranging from 16 to 60; durations of the video ranged from 4 seconds to 1 minute 11 seconds. Videos were taken from the recording of real trials; thus, the dataset has varied lighting, camera angles, and poses.

B. Data Preprocessing

This study involves two types of data: text data for the interview generation process and audio recordings for lie detection. Two distinct preprocessing techniques are applied accordingly.

1) Question generation dataset: Tokenization in the Llama 3 model uses a subword-based approach, enabling the Llama 3 model to efficiently handle words not in the model's vocabulary or the presence of affixes. The input text is converted into ID tokens in tokenization. The ID tokens are then converted into mathematical representations through an embedding layer. Each token ID is mapped to a 3072-dimensional vector according to the embedding size used by Llama 3.

Consequently, an input sequence consisting of n tokens is converted into an embedding matrix of size $n \times 3072$. Then, this matrix is used as the input for the transformer layer in the model. The matrix will then be processed in the transformer layer in the model so that the Llama model can understand the input context.

2) Lie detection dataset: To prepare the audio, sound tracks were separated from the video files using the moviepy library. The extracted audio was then adjusted to meet a common format resampled at 16 kHz to keep the data consistent and manageable for processing. Because the recordings varied in length, each clip was truncated and padded into 30 seconds: shorter clips were extended with silent padding, whereas longer clips were trimmed down.

The response latency data were obtained using the pyAudioAnalysis package through a silence removal process. This method segments the audio by removing non-speech regions using a semi-supervised SVM classifier trained to distinguish between high-energy (speech) and low-energy (non-speech) frames. The resulting speech segments were used to calculate the response latency for the lie detection analysis.

3) Data splitting: For robust model assessment and to reduce the likelihood of overfitting, the dataset was separated into two distinct portions: training and testing. The training subset was used to adjust the model parameters. The testing subset supported hyperparameter tuning and helped monitor overfitting during learning.

In this study, stratified random splitting was applied to maintain the balance between the "truth" and "lie" categories across all subsets. This strategy ensured the preservation of class proportions, preventing bias that might otherwise distort model performance. The split was performed using standard library functions, guaranteeing that the process could be replicated consistently. Data allocation followed a widely accepted distribution in machine learning practice: 75% of the data were used for training and 25% for testing.

C. Building Models

1) Llama 3: The primary methodological challenge in the AQG component was adapting the general-purpose Llama 3 model [28] to the specific domain of financial risk assessment. This was not a simple zero-shot prompting task; it required a specific fine-tuning strategy to achieve the desired domain adaptation. We structured our synthetic dataset as 1) into a Biodata, Analysis, Question format. The model was fine-tuned on this structured data to explicitly learn the task of: a) first analyzing a borrower's financial profile and b) then generating contextually relevant, probing questions [27] based on that analysis. This "analysis-then-question" approach is our key technical adaptation, moving the model from a general text generator to a specialized financial risk interrogator.

To execute this efficiently, we utilized Low-Rank Adaptation (LoRA) for fine-tuning, applying it to the model's attention layers. We leveraged the model's native architecture [29] which includes pre-normalization with RMSNorm [30], Rotary Positional Embeddings [31], and the SwiGLU activation function [32] to ensure stable and efficient convergence during our specialized fine-tuning task.

2) SincNet: We selected SincNet as the audio front-end to process raw waveforms directly. This choice was a critical part of our methodology, as traditional hand-crafted features like MFCCs, which are common in audio processing can suppress narrow-band characteristics such as pitch and formants that are key indicators of deception-related stress [33].

SincNet avoids this by using parametrized sinc-based filters in its first convolutional layer [34]. These filters function as learnable band-pass filters defined by cutoff frequencies f_1 and f_2, which are optimized during our training process. The filter is expressed as Eq. (1):

$$g_w[n, f_1, f_2] = (2f_2 \cdot \mathrm{sinc}(2\pi f_2 n - 2f_1 \cdot \mathrm{sinc}(2\pi f_1 n)) \cdot w[n] \quad (1)$$

where, $w[n]$ is a windowing function applied to smooth the filter edges. By implementing this architecture, our model learns the most meaningful frequency features for lie detection directly from the raw audio, rather than relying on pre-engineered features, which is crucial for our multimodal hypothesis.

3) Multimodal: Multimodal is a single algorithm that combines multiple inputs types of input [35]. This study leverages a multimodal deep learning framework to jointly process two types of input: audio data and response latency. Each modality undergoes preprocessing according to its nature. For audio, the signals are resampled, padded, and truncated to achieve a consistent length and dimensionality. Then, SincNet is used to extract the informative features directly from the raw waveform. The silent segments in the audio were identified and removed to capture the response latency, and the resulting features were processed through fully connected layers to produce meaningful numerical representations.

The feature outputs from both modalities are concatenated into a single vector during fusion, forming a shared representation. This combined embedding is then refined through additional fully connected layers to integrate cross-modal information. The final classification layer predicts whether the response reflects truthfulness or deception.

D. Experimental Design

Llama 3 (version 3.2) with 3 billion parameters was employed for the automated question generation task. Fine-tuning was conducted using a learning rate of 2×10^{-4}, over 10 epochs, with a batch size of 3 applied to training and evaluation. During the generation phase, the decoding process was configured with a maximum sequence length of 512 tokens, Top-K sampling set to 50, a temperature of 0.6, and Top-P of 0.95. A multimodal model was trained across 50 epochs with a batch size of 8 for deception detection. The system takes two inputs: raw audio and response latency. Audio signals are processed through SincNet to extract features, whereas response latency is transformed using fully connected layers. Then, the outputs from both pathways were fused for joint representation and final classification.

IV. Experiments and Results

In this section, we explain the results of our research and provide a comprehensive discussion.

A. Evaluation Parameters

1) Question generator evaluation: In this case, the effectiveness of the LLM Llama 3 in generating questions was assessed using perplexity. As described in [36], perplexity quantifies how well a language model or probability distribution predicts a given sample. Lower perplexity values reflect stronger performance, indicating that the model is more confident and precise in forecasting the next token in a sequence. The mathematical formulation of perplexity is expressed as Eq. (2):

$$perplexity = 2^{-\frac{1}{M}\sum_{j=1}^{M}\log_2 \hat{P}(y_j)} \quad (2)$$

2) Lie detection model evaluation: The multimodal deep learning-based lie detection model was assessed using conventional classification metrics derived from the confusion matrix (TABLE III The confusion matrix offers a detailed breakdown of the prediction outcomes by distinguishing between true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) [37]. In the context of this binary classification task (Honest vs. Deceptive), the matrix is organized as follows:

Based on [38], the following metrics were calculated from the confusion matrix:

- **Accuracy** reflects the proportion of correctly classified samples relative to the total number of cases and serves as an overall indicator of correctness. The equation is shown as Eq. (3):

$$\text{Accuracy} = \frac{TP + TN}{Total\ Number\ of\ Cases}. \quad (3)$$

- **Precision** denotes the fraction of TPs within all predicted positives, highlighting the effectiveness of the model in minimizing FPs. The equation is shown as Eq. (4):

$$Precision = \frac{TP}{TP\ +\ FP}. \quad (4)$$

- **Recall (or Sensitivity)** measures the share of TPs among all actual positives, illustrating the ability of the model to capture all relevant instances. The equation is shown as Eq. (5):

$$Recall = \frac{TP}{TP\ +\ FN}. \quad (5)$$

- The **F1-score**, defined as the harmonic mean of Precision and Recall, offers a balanced metric that is especially valuable in cases of class imbalance. The equations is shown as Eq. (6):

$$F1 - score = \frac{2 \times p \times r}{p + r} \quad (6)$$

- **Specificity** quantifies the proportion of TNs within all actual negatives, indicating the strength of the model in avoiding false alarms. The equation is shown as Eq. (7):

$$Specificty = \frac{TN}{TN + FP} \quad (7)$$

Together, these metrics provide a comprehensive evaluation framework for assessing the question generation component and the lie detection performance of the proposed system.

TABLE III. Confusion Matrix of Two Classes

		Prediction Label	
		Class 0	**Class 1**
Ground Truth	Class 0	TN	FP
	Class 1	FN	TP

B. Hardware and Software Specification

All experiments in this study were executed using a high-performance computing infrastructure. Table IV shows the hardware and software specifications used for training the LLM and the multimodal deep learning models.

TABLE IV. Hardware and Software Specifications

Hardware	
GPU	RTX 4090
VRAM	24 GB
Ram	128 GB
Storage	1 TB
Software	
Programming Software	VSCode
Programming Language	Python 3.10

C. Performance Analysis

1) Question generator performance: Fig. 4 presents the progression of loss values for training and testing datasets over the course of epochs. During the initial epochs (0–2), training and testing losses sharply declined, indicating that the model quickly captured fundamental patterns and exhibited strong generalization. Between epochs 3 and 6, the training loss continues to decrease. However, after epoch 6, the testing loss levels off and subsequently begins to rise, signaling the

emergence of overfitting. This outcome suggests that the model's large capacity led it to memorize the training data rather than extract broadly applicable features. Approaches such as early stopping and applying regularization methods should be considered to address this issue and preserve generalization.

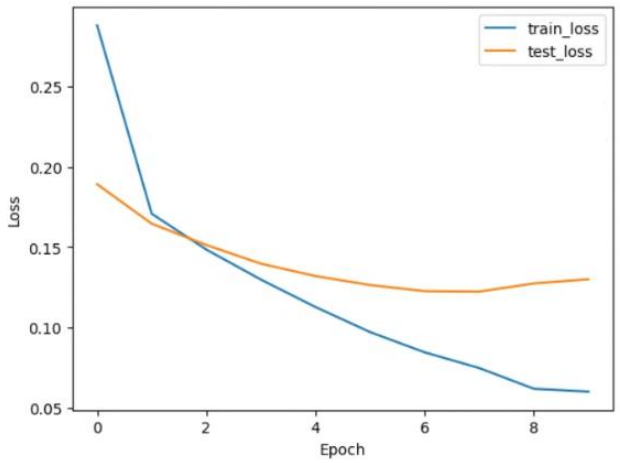

Fig. 4. Changes in the LLM train and test loss with respect to epoch.

The perplexity value for the trained model ranged from 2 to 3, indicating a good predictive capability for subsequent tokens within the language corpus. A low perplexity score in LLMs, such as Llama 3, indicates efficient modeling of the language probability distribution and more accurate predictions. While low perplexity is a positive indicator of performance, the observed overfitting in the test loss highlights the importance of maintaining model generalization for real-world data application.

TABLE Vpresents an example of the output generated by the Llama 3 model. The analysis demonstrates accuracy and professionalism by incorporating details such as estimated monthly installments, their ratio to income, and consideration of rental status as a non-financial risk factor. The generated questions are contextually relevant and exploratory, addressing key aspects such as payment history, emergency fund availability, and preparedness for unforeseen expenses. These results highlight the model's strong potential for producing meaningful analyses and credit-related inquiries, suggesting that it can be effectively integrated into an automated credit scoring system with further refinement.

2) Lie detection performance: This section presents the performance of the lie detection model on the training and testing datasets. The model was trained for 50 epochs with a batch size of 8. Fig. 5 illustrates the evolution of loss values across epochs for both datasets, providing insights into the learning behavior of the model during training and its generalizability to unseen data.

The graph illustrates the progression of the loss values throughout the training of the deep learning model. The horizontal axis represents the number of epochs, whereas the vertical axis shows the loss value. Two curves are displayed: the blue curve corresponds to the training loss and the orange curve represents the testing loss. Both losses are relatively high (approximately 0.7). As training progressed, the training loss dropped steeply, reaching approximately 0.05, reflecting effective learning on the training data. The testing loss also decreased but in a gradual and stable manner, plateauing between epochs 20 and 35 before showing a slight increase toward the end of training. This trend suggests the onset of overfitting after epoch 35, when the model becomes too tailored to the training data, resulting in reduced performance on unseen data.

TABLE V. EXAMPLE OF GENERATED OUTPUT FROM THE TRAINED MODEL

Biodata	Analysis	Questions
Full Name: Rudi Hartono; **Province**: West Java; **City**: Bekasi; **Educational Background**: High School; **Marital Status**: Single; **Reason for Loan**: Household needs; **Residence**: Rental; **Personal Vehicle**: Motorcycle; **Company Name**: PT. Sentosa Abadi; **Occupation**: Private employee; **Industry**: Industry/Factory; **Position**: Supervisor; **Net Monthly Income**: Rp 8,500,000; **Family Monthly Income**: Rp 8,500,000; **Length of Employment**: 4 years; **Emergency Contact**: Parent; **Emergency Contact Name**: Sri Hartono; **Loan Amount**: Rp 15,000,000; **Loan Tenure (Months)**: 18 months; **Loan Interest (Per Month)**: 6.0%	With a loan amount of Rp. 15,000,000 and a tenor of 18 months, the estimated monthly installment is approximately Rp. 1,733,333.33 (with 6% monthly interest). The total repayment will reach Rp. 31,200,000, meaning that around 9.7% of the borrower's monthly income will be allocated for loan repayment.	1. Have you ever been late in paying installments or loans before? 2. What is your plan to ensure that this installment will not interfere with your routine expenses? 3. Do you currently have any other ongoing loans or credits? If yes, what is the total monthly installment you need to pay?

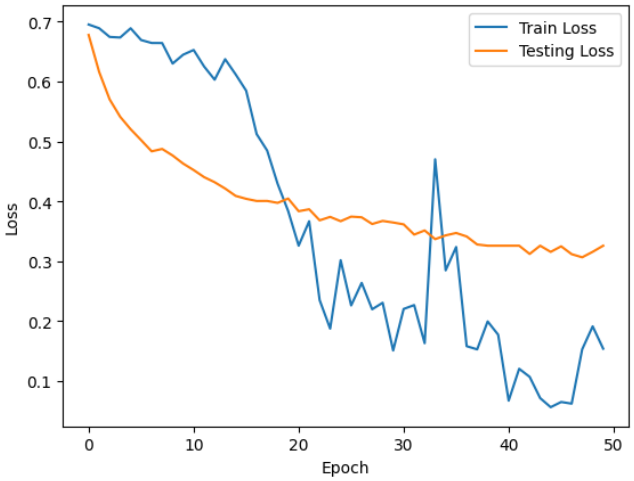

Fig. 5. Change in the lie detection model train and test loss with respect to epoch.

Fig. 6 depicts the evolution of the model's accuracy during the 50-epoch training process, with the horizontal axis representing epochs and the vertical axis showing the accuracy

in percentage. There are two curves: the blue curve represents the training accuracy, whereas the orange curve shows the testing accuracy. At the beginning of training, the training accuracy started at a relatively low and fluctuating value, but generally increased sharply to over 90%. Despite this increase, the training accuracy curve shows considerable fluctuations at several points, likely owing to unstable training data or varying batches.

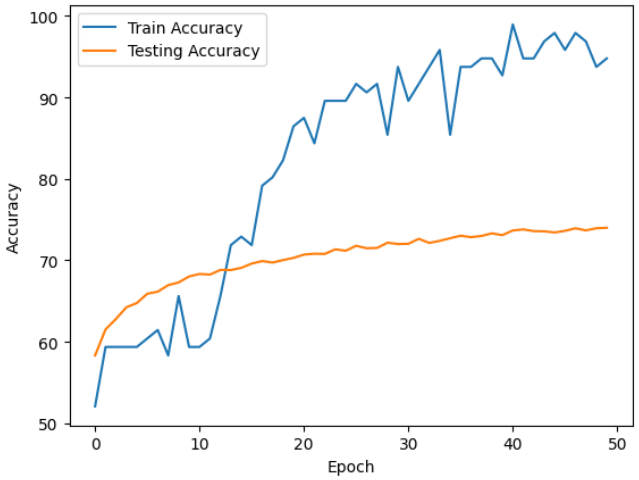

Fig. 6. Change in the lie detection model train and test accuracy with respect to epoch.

Conversely, the testing accuracy curve appeared smoother and more consistent, showing a gradual increase from approximately 58% to nearly 70%, before stabilizing toward the end of training. The substantial gap between the training and testing accuracy curves highlights the presence of significant overfitting. This discrepancy indicates that although the model achieved outstanding performance on the training data, it struggled to effectively generalize to unseen data.

The developed lie detection model achieved 100% accuracy on the training set (TABLE VIand TABLE VIIThis outcome shows that the model successfully captured the underlying patterns of the training data. However, when evaluated on the test set, which contained an evenly balanced distribution of the "honest" (class 0) and "deceptive" (class 1) classes, its accuracy dropped to 70% (TABLE VI. This decline in performance further confirms the occurrence of overfitting, indicating that while the model excelled in memorizing the training data, it exhibited limited generalization capability when applied to new, unseen samples.

TABLE VI. CONFUSION MATRIX OF THE LIE DETECTION MODEL ON TRAINING DATA

	Class	Prediction	
		Honest	Lie
Ground Truth	Honest	45	0
	Lie	0	46

TABLE VII. PERFORMANCE OF THE LIE DETECTION MODEL ON TRAINING DATA

	Accuracy	Precision	Recall	F1 Score	Specificity
Train	100%	100%	100%	100%	100%

Based on the confusion matrix in TABLE VIIIand TABLE IXfor the testing data, 10 honest samples were correctly classified and 5 honest samples were misclassified as deceptive (FP). Of the 15 deceptive samples, 11 were correctly classified, whereas 4 others were misidentified as honest (FN). The precision and recall values for the deceptive class were 68.8% and 73.3%, respectively. This relatively small difference indicates that the model has a balanced performance in detecting deception, both in terms of its ability to correctly identify deceptive cases (recall) and the accuracy of predicting deception (precision). However, the slightly lower precision compared with recall suggests that there are still some incorrect deceptive predictions (FPs), although the overall performance of the model can be considered good in detecting deception. Specificity, the ability of the model to correctly detect honesty, was 66.67%.

TABLE VIII. CONFUSION MATRIX OF THE LIE DETECTION MODEL ON TESTING DATA

	Class	Prediction	
		Honest	Lie
Ground Truth	Honest	10	5
	Lie	5	11

TABLE IX. PERFORMANCE OF THE LIE DETECTION MODEL ON TESTING DATA

	Accuracy	Precision	Recall	F1 Score	Specificity
Train	70%	68.8%	73.3%	70.9%	66.67%

In a real-world context, misclassifying an honest person as deceptive can have a negative impact on reputation and trust. Conversely, if a deceptive person goes undetected as honest, other forms of loss can occur. Therefore, a better balance between the model's sensitivity (recall) and specificity is required to improve its overall performance.

D. Discussion

The experimental results of this study demonstrate the potential and current limitations of employing artificial intelligence to improve verification processes in online lending, particularly through automated lie detection. The proposed system, which integrates a LLM for personalized question generation and a multimodal deep learning model for deception detection, represents a novel approach to reducing the risk of loan default.

The LLM-based question generator, built upon Llama 3, successfully generated contextually relevant and individualized questions by analyzing borrower biodata. The observed perplexity scores between 2 and 3 indicate that the model learned meaningful representations of the input data and could effectively predict token sequences. Nonetheless, the training dynamics revealed signs of overfitting,

particularly beyond epoch 6, where the testing loss began to increase despite continuous improvement in the training loss. This phenomenon suggests that the model began to memorize training samples rather than learning generalized patterns. Future implementations should consider employing regularization techniques such as dropout as well as strategies such as early stopping and k-fold cross-validation to mitigate this issue.

The lie detection model demonstrated strong learning ability during training, achieving 100% accuracy. However, the performance on the testing dataset declined to 70% accuracy, with additional metrics including a precision of 68.8%, recall of 73.3%, F1 score of 70.9%, and specificity of 66.67%. These figures reflect a moderate level of generalization and indicate that the model was more proficient in detecting deceptive behavior (as shown by its relatively higher recall) than in avoiding FPs. This trade-off between sensitivity and specificity is particularly important in financial contexts: while failing to detect deception may result in financial loss, false identification of honest applicants can damage user trust and accessibility.

The discrepancy between the training and testing performance, along with the presence of FPs and FNs, indicates the need for further refinement of the model. The domain gap between the courtroom trial audio data used for training and the real-world voice patterns of loan applicants may be contributing factors. Therefore, the collection and use of domain-specific audio data, sourced directly from online lending environments, is recommended to improve model adaptability and contextual accuracy.

Moreover, while the technical feasibility of the proposed system is evident, its practical deployment raises several ethical concerns. The automated classification of debtor honesty based on vocal features and response timing may introduce biases, especially for individuals with speech disorders, anxiety, or cultural-linguistic variations. These risks necessitate the implementation of fairness-aware machine learning techniques and the potential integration of a human-in-the-loop review mechanism.

In summary, the results confirm that the integration of adaptive LLM-based questioning with voice and response latency analysis holds substantial promise for improving the objectivity and robustness of digital credit verification systems. However, to transition this system from proof-of-concept to deployment, further work is required in terms of data expansion, overfitting mitigation, domain adaptation, and ethical governance. This study's methodology provides a solid foundation for the future development of scalable, intelligent decision-support tools for financial technology platforms.

V. Conclusion

This study presents an AI-driven lie detection system tailored for online lending applications, integrating a LLM with a multimodal deep learning framework. The proposed system aims to improve the borrower verification process by generating personalized, context-aware interview questions using Llama 3 and subsequently evaluating applicant honesty through vocal acoustic features and response latency.

The experimental results show that the system performs well during the training phases, achieving 100% accuracy in lie detection on the training data and producing highly relevant, individualized questions. However, generalization remains a challenge, as evidenced by the drop in testing accuracy to 70% and indications of overfitting in the question generation and deception detection components. Despite this, the model achieved a balanced performance on unseen data, with an F1 score of 70.9%, precision of 68.8%, recall of 73.3%, and specificity of 66.67%.

This integrated framework represents a significant advancement over existing methods. Unlike traditional, static questionnaires, the LLM-driven interview prevents rehearsed answers; unlike rule-based scoring, it adds behavioral analysis; and unlike unimodal models, its fusion of vocal acoustics and latency provides a more robust deception signal. This scalable framework thus shows strong potential as a decision-support tool. Its automation enables thousands of simultaneous interviews, reducing labor costs and financial losses from NPLs, with broader implications for recruitment, insurance, and fraud investigations.

Future research should focus on mitigating overfitting through hyperparameter tuning, applying cross-validation techniques, expanding and diversifying the dataset, and incorporating advanced feature extraction methods. Furthermore, ethical considerations, including bias mitigation and human oversight, must be integrated into the deployment pipeline to ensure fairness and transparency in real-world applications.

Acknowledgment

This research was supported by the Program Penelitian Tahun Anggaran 2025, under the Directorate of Research and Community Service, Directorate General of Research and Development (DIKTI), Ministry of Higher Education, Science, and Technology of Indonesia, through Research Grant No. PKS-507/UN2.RST/HKP.05.00/2025. The authors also acknowledge the additional support provided by PT. Global Risk Management (GRM), PT. Seleris Meditekno Internasional, Data Science Center (DSC), and the Laboratory of Bioinformatics and Advanced Computing, Department of Mathematics, Universitas Indonesia for the successful completion of this research.

References

[1] Asosiasi Penyedia Jasa Internet Indonesia, "Hasil survei internet APJII.", *Online*, Nov. 2024.

[2] D. Wati and T. Syahfitri, "Dampak pinjaman online bagi masyarakat," *Community Dev. J.*, vol. 2, no. 3, pp. 1181–1186, Nov. 2022.

[3] J. Z. Y. Arvante, "Dampak permasalahan pinjaman online dan perlindungan hukum bagi konsumen pinjaman online," *Ikatan Penulis Mahasiswa Hukum Indones. Law J.*, vol. 2, no. 1, pp. 73–87, Feb. 2022.

[4] R. Kartika and M. Umam, "Tingkat wanprestasi 90 peer to peer lending selama pandemi COVID-19 di Indonesia," *Akuntabilitas J. Ilm. Ilmu-Ilmu Ekon.*, vol. 14, no. 1, pp. 31–40, Nov. 2021.

[5] A. Ahdiat, "Kredit macet pinjol tembus Rp1,9 triliun akhir semester I 2024," *Katadata Databoks*. Nov. 2024.

[6] C. M. Annur, "Nilai kredit macet pinjol berdasarkan kelompok usia penerima pinjaman," *Katadata Databoks*. Aug. 2023.

[7] P. F. Andini, I. Zidane, H. D. Wahyuni, M. Y. I. Rabbani, and L. N. Yuliati, "Perlindungan konsumen pinjaman online yang mengalami

keterlambatan pembayaran cicilan," *Policy Brief Pertan. Kelautan Biosains Trop.*, vol. 4, no. 4, pp. 381–386, Feb. 2022.

[8] U. S. Liman, "AFPI memanfaatkan kecerdasan buatan cegah penipuan pinjaman daring," *Antara News*, May 6, 2025.

[9] M. H. Simanjuntak, "OJK: Literasi keuangan digital cegah masyarakat dari pinjol ilegal," *Antara News*, Nov. 4, 2024.

[10] M. Suhayati, "FENOMENA KREDIT MACET PINJAMAN ONLINE," *Isu Sepekan Bidang Ekkuinbang, Komisi XI,* Jul. 2023.

[11] F. C. Akyon, D. Cavusoglu, C. Cengiz, S. O. Altinuc, and A. Temizel, "Automated question generation and question answering from Turkish texts using text-to-text transformers," *arXiv preprint,* Nov. 2021.

[12] S. Bulathwela, H. Muse, and E. Yilmaz, "Scalable educational question generation with pre-trained language models," *arXiv preprint,* Apr. 2022.

[13] G. H. Alferez, "Comprehensive question and answer generation with LLaMA 2," *M.S. thesis*, Southern Adventist Univ., Tennessee, 2024.

[14] P. Ekman, M. O'Sullivan, W. V. Friesen, and K. R. Scherer, "Face, voice, and body in detecting deceit," *J. Nonverbal Behav.*, vol. 15, no. 2, pp. 125–135, Jun. 1991.

[15] C. Kirchhuebel, "The acoustic and temporal characteristics of deceptive speech," *Ph.D. dissertation*, Univ. of York, York, U.K., Oct. 2013.

[16] B. DePaulo, J. J. Lindsay, B. E. Malone, L. Muhlenbruck, K. Charlton, and H. Cooper, "Cues to deception," *Psychol. Bull.*, vol. 129, no. 1, pp. 74–118, Jan. 2003.

[17] R. Loconte, R. Russo, P. Capuozzo, P. Pietrini, and G. Sartori, "Verbal lie detection using large language models," *Sci. Rep.*, vol. 13, no. 1, p. 22849, Dec. 2023.

[18] J. Feng, "DeepLie: Detect lies with facial expression (computer vision)," 2021.

[19] R. Rill-García, H. J. Escalante, L. Villaseñor-Pineda, and V. Reyes-Meza, "High-level features for multimodal deception detection in videos," *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, pp. 1565–1573, Jun. 2019.

[20] Z. Wu, B. Singh, L. Davis, and V. Subrahmanian, "Deception detection in videos", *AAAI*, vol. 32, no. 1, p. 11502, Apr. 2018.

[21] H. Rohde and P. A. Finkelstein, "An acoustic automated lie detector," *Semant. Scholar Corpus*, Jan. 2019.

[22] F. M. Talaat, "Explainable enhanced recurrent neural network for lie detection using voice stress analysis," *Multimed. Tools Appl.*, vol. 83, no. 11, pp. 32277–32299, Nov. 2024.

[23] G. Krishnamurthy, N. Majumder, S. Poria, and E. Cambria, "A deep learning approach for multimodal deception detection," *arXiv preprint*, Mar. 2018.

[24] H. U. D. Ahmed Khan, U. I. Bajwa, N. I. Ratyal, F. Zhang, and M. W. Anwar, "Deception detection in videos using the facial action coding system," *Multimed. Tools Appl.*, vol. 84, no. 9, pp. 6429–6443, Apr. 2024.

[25] G. Melis, M. Ursino, C. Scarpazza, A. Zangrossi, and G. Sartori, "Detecting lies in investigative interviews through the analysis of response latencies and error rates to unexpected questions," *Sci. Rep.*, vol. 14, no. 1, p. 12268, May. 2024.

[26] V. Pérez-Rosas, M. Abouelenien, R. Mihalcea, and M. Burzo, "Deception detection using real-life trial data," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI '15)*. pp. 59–66, Nov. 2015,

[27] M. Ridho, A. Bustamam, and R. Adnan, "Reconstruction of the Phi-2 Method for Question-Answering Related to Diabetes Disease Using the MedAlpaca Dataset," Jambura Journal of Biomathematics (JJBM), vol. 6, no. 3, pp. 183–187, 2025

[28] A. Dubey *et al.*, "The LLaMA 3 herd of models,", *arXiv preprint*. Nov. 2024.

[29] M. Tamang, "Build your own Llama 3 architecture from scratch using PyTorch," *Towards AI*. Sep. 2024.

[30] B. Zhang and R. Sennrich, "Root mean square layer normalization," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst. (NeurIPS).* 2019.

[31] J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu, "RoFormer: Enhanced transformer with rotary position embedding," *Neurocomputing*, vol. 568, p. 127063, Feb. 2024.

[32] N. Shazeer, "GLU Variants Improve Transformer," *arXiv preprint,* Feb 2020.

[33] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," in *Proc. Interspeech*, Aug. 2017.

[34] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with SincNet," *2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 1021–1028, Jul. 2018.

[35] N. Hilmizen, A. Bustamam and D. Sarwinda, "The Multimodal Deep Learning for Diagnosing COVID-19 Pneumonia from Chest CT-Scan and X-Ray Images," *2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, Yogyakarta, Indonesia, pp. 26-31, 2020.

[36] T. B. Brown *et al.*, "Language models are few-shot learners," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 1877–1901, May. 2020.

[37] J. Patterson and A. Gibson, *Deep learning: A practitioner's approach*, 1st ed. Sebastopol, CA, USA: O'Reilly Media, Inc., 2017.

[38] M. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," *Int. J. Data Min. Knowl. Manag. Process*, vol. 5, pp. 1–11, Mar. 2015.