Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/180609

Analysis of Deception Detection Databases Using Mathematical Statistics

Aikumis Omirali, Rakhima Zhumaliyeva, Didar Bayazitov, and Kanat Kozhakhmet*

School of Digital Technologies, Narxoz University, Almaty, Kazakhstan

Received: 12 May 2024, Revised: 22 Jun. 2024, Accepted: 21 Aug. 2024
Published online: 1 Nov. 2024

Abstract: With the rise of digital media platforms and social networks, distinguishing between trustworthy and false information has become increasingly complicated. This has posed challenges in shaping public opinion and making informed decisions. In order to tackle this issue, this research paper presents a novel dataset based on experimental findings, which is specifically designed for detecting factual and fabricated statements in video content. The dataset was carefully curated through the production of independent videos in diverse scenarios, capturing both honest and deceitful contexts. The paper provides a thorough description of the methodologies used in collecting, processing, extracting features, and annotating the data, highlighting its credibility and representativeness. In addition, the paper offers a comprehensive analysis of existing databases in the deception detection tasks, underscoring the significance of this new dataset.

Keywords: Deception detection, True and false statements, Dataset creation, Video materials, Data annotation, Feature extraction, Machine learning models, Social networks, Digital media, Information authenticity

1 Introduction

In recent decades, the information space has undergone significant changes, leading to various problems associated with the dissemination of true information. Due to the rapid development of social networks and digital media platforms, it has become difficult to distinguish truthful information from false statements. This causes not only embarrassment and distrust among users, but can also have a negative impact on public opinion, decision-making and the formation of political preferences [1].

In the context of such challenges, it becomes important to develop effective methods for automatic detection of true and (or) false statements. For this purpose it is necessary to have access to data sets that will be used for training and evaluating machine learning models.

In this research paper, we present an experimentally created data set specially designed for the task of detecting true and (or) false statements. Our data set is based on video materials shot by us independently using various scenarios and conditions. This approach allows us to obtain more realistic data that reflects a variety of situations in which true and false statements may occur.

We also describe the methods we used to process the collected data. These methods include both techniques for preprocessing video materials and algorithms for feature extraction, which make it possible to efficiently present data for subsequent analysis and training of models. Particular attention is paid to data markup and annotation methods to create reliable and accurate labels for each video fragment in the dataset.

Our work also includes a description of the data collection method. We describe in detail the process of shooting video and the parameters that we have set to create a variety of scenarios. This aspect of the study is important to ensure the validity and representativeness of our dataset.

As a result of our work, we provide the scientific community with a valuable set of data for the task of detecting true and (or) false statements.

2 Analysis of Existing Databases in the Field of Research

Up to the present day, the identification of deception remains a pressing challenge within the realm of artificial

* Corresponding author e-mail: rakhimazhumaliyeva@gmail.com

intelligence (AI). Dedicated deep learning and machine learning engineers are actively engaged in crafting algorithms tailored to address this task. Within the context of building and evaluating various models, the use of distinct datasets is deemed crucial. In the domain of scam detection, the availability of public datasets is rather limited. In this section, we present an overview of eight databases designed for the purpose of deception detection, accompanied by comprehensive descriptions of each.

2.1 "Box of Lies" Dataset

The "Box of Lies" dataset [2] comprises 25 publicly accessible videos from the "Box of Deception" series, involving 26 participants. Each video, with an average duration of 6 minutes, includes approximately 3 rounds of deception. This dataset contains a total of 68 rounds, evenly split between truthful and deceptive instances. It serves as a valuable resource for understanding real-life deception behaviors in a controlled setting.

2.2 Multimodal Dataset for Deception Detection

This dataset [3] encompasses acoustic, visual, physiological, and thermal data modalities, involving 30 respondents. The use of diverse sensors, webcams, and a thermal camera provides rich and multi-dimensional data for deception analysis. It is especially suitable for exploring the correlation between physiological responses and deceptive behaviors.

2.3 "Bag of Lies": Deception Detection Database

With data modalities including audio, video, EEG, and Eye Gaze, this dataset [4] comprises 325 annotated records from 35 participants. The inclusion of EEG and Eye Gaze data adds a neurophysiological dimension to deception detection research. Additionally, the presence of informative images enhances the experimental design.

2.4 Database with a High Probability of Detecting Facial Deception

The database [5] assembled by Lin Su and Martin D. Levin includes 324 videos with a deliberate inconsistency in lighting conditions and backgrounds. This variation challenges deception detection models to handle real-world scenarios. The dataset offers an opportunity to assess the robustness of algorithms under diverse conditions.

2.5 Miami University Database

The Miami University Deception Detection Database [6] houses 320 videos from 80 participants, representing various demographic characteristics and statement valence. The dataset's diversity is advantageous for training and evaluating AI models across different contexts, making it valuable for real-world applications.

2.6 "Silesian" Database

Comprising 101 videos with over 1.1 million frames, the "Silesian" database [7] focuses on the interaction between truth-telling and lying. The controlled research center environment and high-speed camera contribute to a detailed analysis of subtle indicators of deception, making it suitable for fine-grained studies.

2.7 Real-life Trial Data

This dataset [8] consists of 121 videos obtained from court hearings, police interrogations, and the Innocence Project website. The inclusion of real-life scenarios provides insights into deception detection in authentic, high-stakes situations. The dataset's diversity in age and context enhances its practical relevance.

2.8 UR Dataset

The UR Lying Dataset [9] offers 107 videos with reliable camera conditions and minimal background interference. Its focus on controlled recording conditions provides a baseline for evaluating deception detection algorithms in a controlled environment.

Table 1 provides an at-a-glance overview of the key characteristics of each deception detection database. The availability of diverse datasets is paramount for advancing deception detection research in artificial intelligence. Furthermore, the combination of datasets and cross-dataset evaluations may yield insights into the generalization capabilities of AI-based deception detection systems. These datasets collectively contribute to the ongoing pursuit of effective and reliable deception detection in AI applications.

3 Methodology

In this section, we outline the process of creating a multimodal dataset for deception detection. To achieve this goal, we conducted a series of experiments in which participants were asked to provide deceptive or truthful responses to different types of questions, including misleading questions, trap questions, and straightforward

Table 1: Summary of Deception Detection Databases

Database Name	Data Modalities	Participants	No of Videos	Average Video Duration	Unique Features
”Box of Lies” Dataset	Video	26	25	6 minutes	Controlled setting with real-life deception scenarios.
Multimodal Dataset for Deception Detection	Acoustic, Visual, Physiological, Thermal	30	N/A	N/A	Diverse multimodal data for comprehensive analysis.
”Bag of Lies”: Deception Detection Database	Audio, Video, EEG, Eye Gaze	35	325	3 to 42 seconds	Inclusion of EEG and Eye Gaze data.
Database with High Probability of Detecting Facial Deception	Video	N/A	324	20 seconds	Varied lighting and background conditions.
Miami University Database	Video, Audio	80	320	N/A	Diversity in demographic characteristics and valence.
”Silesian” Database	Video	N/A	320	N/A	High-speed camera in a controlled environment.
Real-life Trial Data	Video	56	121	28 seconds	Real-life scenarios, diverse age range, and contexts.
UR Lying Dataset	Video	29	107	23 minutes	Controlled recording conditions.

questions. The selection of topics was motivated by previous research where linguistic behavior was studied under similar conditions, revealing significant distinctions between truth-tellers and deceivers [10]-[13].

3.1 Materials

The recordings are recorded using the” Sony 4K Handycam ” balanced SteadyShot optical image stabilizer and camera with 20x optical zoom with 4K (3840 x 2160) video definition and 60 FPS. SteadyShot is Sony’s unique balanced optical image stabilizer with a suspension that is about 13 times more efficient than other optical systems. The AX43 is equipped with a special lens that provides 4K quality at all focal lengths from wide-angle to telephoto.

The Ax43’s three-capsule microphone registers sound in five directions. Combined with advanced audio processing, this greatly improves the quality of audio recording. The noise level is reduced by about 40%, and in a stereo or 5.1 - channel signal (AVCHD), individual channels sound clearer, which makes the dialogue clearer. Other technical characteristics of the device are shown in Table 2.

3.2 Participants

20 men and 10 women participated in the study. The age range of respondents was 17-19 years. Such a choice of participants allowed us to obtain data from representatives of the younger generation and evaluate the features of their emotional and verbal reactions to questions.

3.3 Procedure

The process involved data collection through video recordings, with the assistance of respondents participating in interviews. To collect the data used in this study, video recordings were made with the help of respondents. The source of the videos were the control participants who were invited to participate in interviews. Data collection process includes following steps:

1.Respondents and groups

A total of 30 respondents were involved, who were divided into 3 groups of 10 people each. Each group presented a certain condition for answering questions: honest answers, false answers with the help of guidance and free answers. This allowed us to obtain data reflecting different situations and approaches to statements.

Table 2: Technical characteristics of ”Sony 4K Handycam”

Description name	Value
Effective pixels in video capture	about 8.29 MP (16: 9)
Matrix type	1/2 inch (7.20 mm) CMOS Exmor R
Lens type, filter diameter	ZEISS Vario – Sonnar T, 55 mm
Focus	Diagram: F2. 0-3. 8, Focus distance: f=4.4-88mm, f=26.8-536.0 mm (16:9), f = 1 1 / 16-21 1 / 8 inches
Audio	Built-in zoom microphone, Audio recording format: LPCM 2-Kan. (48 kHz / 16 bit), Dolby Digital 5.1, Dolby Digital 5.1 Creator, Dolby Digital 2-Kan. Stereo, Dolby Digital Stereo Creator, MPEG-4 AAC-LC 2-Kan.

2.Questions and categories
The questions were pre-prepared and divided into different categories. It is important to note that trap questions were included, which were questions asked from ambush, forcing respondents to think and causing a vivid emotional reaction. This made it possible to record the facial expressions and movements of the participants and study them in the context of the truth or falsity of statements.
3.Duration of the interview
Each interview lasted 120 minutes, during which the respondents answered questions and demonstrated their facial expressions and movements. The length of the interview allowed us to obtain enough data to analyze and evaluate the performance of models.
4.Data aggregation
From the video recording process, 81 videos were shot, which were cut and compiled according to the answer to each question.

As a result, an ”experimental data set” was created, consisting of 170 false statements, 202 true statements, a total of 372 video recordings. The average duration of each passage is 5-10 seconds.

Video recordings were made in an experimental environment, in an audience covered in good light.

Having publicly available data sets to solve a particular problem (i.e., in this case, lie detection) is essential to encourage and accelerate progress in solving the relevant problem. In a sense, this approach is an example of open innovation, although it benefits extensive correspondence (including universities and companies) and improves its direct impact [8] compared to a set of individually developed and stored data (i.e., can be described as closed innovation). Although there are several jobs that perform lie detection tasks, the existing data sets are considered very small.

Thus, using the described data collection methods, video recordings were obtained with the participation of respondents, reflecting their answers to questions and demonstrating their facial expressions and movements. These data are the basis for the analysis and development of models for detecting true and (or) false statements.

3.4 Video preprocessing

After the shooting was completed, we preprocessed the video clips in order to improve the quality and prepare the data for further analysis. This included image stabilization, noise reduction, and other techniques to ensure uniformity and clarity of video clips.

Data preprocessing plays a key role in preparing raw video data for analysis. In this code, the dataset consists of misleading and truthful images. In the context of deception detection, efficient data preprocessing and tagging are important steps to ensure that the input data is in a model-appropriate format and contains relevant information for deception detection [4]. Data preprocessing steps include audio, video and text data preprocessing.

Audio download and preprocessing. Audio files are extracted through the *Librosa* library. Audio data is downloaded from video files, converted to a waveform, and then converted to midrange kepstral coefficients (MFCC) using the *librosa.feature.mfcc* function. This step provides the presentation of audio content in the video.

Sound features. The audio modality gives an idea of the speech patterns and vocal signals associated with cheating. To extract audio files, the following steps are followed:

a)convert video to audio. Video clips first converted into audio files using the videofileclip class from *moviepy.editor* module. This step will allow us to extract audio signals from the video.
b)audio download. The librosa library is used to download audio files and obtain audio data and sampling frequency. The *librosa.load* function reads the audio file and returns the audio data as the time series and the sample frequency at which the sound was recorded.

c)MFCC extraction. Mel-frequency cepstrum (MFCC) are widely used audio functions that capture the spectral characteristics of an audio signal. The *librosa.feature.mfcc* function is applied to audio data using sampling frequency as a parameter to output mfcc functions. The MFCC is obtained by dividing the audio signal by the frequency range and calculating the kepstral coefficients.

Download and pre-edit text. Text objects are extracted using the PyTesseract library. In a set of real-life data, the code for each frame in the video converts the frame to grayscale, performs Optical Character Recognition (OCR) using pytesseract, and combines the resulting text. This step captures the textual information in the video frames. Text descriptions from the ”experimental data set” were separated by the ”Speech-to-Text API” provided by Google. The speech to text conversion API can recognize more than 125 languages and their variants, for this reason, it was used to obtain text descriptions of video recordings in Kazakh and Russian.

The text from each frame is combined to form a single text view for the entire video. This summary text view serves as an input to the text branch of the model.

Download and pre-edit video. Video elements are extracted using the dlib library. The code uses a face detector to detect faces in each frame of the video. It then uses a form predictor to get page orientations. These facial orientations represent the features of the image and give an idea of facial expressions and gestures.

Video features. The video modality captures facial expressions, head movements and other visual signals related to cheating [13]. To extract video snippets, the following steps are performed:

a)face recognition. The dlib library is used to identify faces in each video frame. The *dlib.get_frontal_face_detector* function returns a face detector object that can detect faces in an image or video frame [14].
b)definition of the reference point. After the faces are defined the *dlib.shape_predictor* function is used to predict page orientations. The *shape_predictor_68_face_landmarks.dat* markup file contains pre-prepared models for predicting the 68 landmarks on the page. Dlib 68 point face orientation detection is a computer vision algorithm that detects 68 face orientations such as eyes, nose and mouth.
c)getting landmarks. For each identified person, the coordinates of the 68 page landmarks are taken and stored in the form of video segments. These reference coordinates provide valuable information about facial expressions and movement.

The figure **??** shows an example of a 68-point Dlib model. There we see points from 1 to 68.

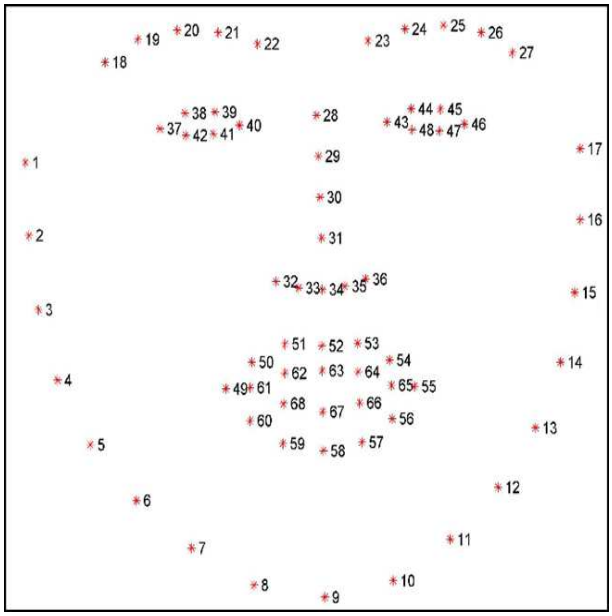

Fig. 1: 68-point dlib page functions

3.5 Data markup

After preliminary processing of the video materials, we carried out data markup in order to assign truth labels to statements in each video fragment. We have developed an annotation system that includes experienced animators who evaluated each statement based on its compliance with actual events and the reliability of the information. The data markup was carried out taking into account strict criteria and internal verification to ensure the accuracy and reliability of the received labels.

Feature Extraction. To effectively represent the data and create the basis for training machine learning models, we applied feature extraction methods. These methods included optical flow analysis, texture characteristics extraction, and inter-frame comparison to capture important aspects of video fragments related to the truth or falsity of statements.

Formation of a data set. Finally, based on pre-processed video materials, posted data and extracted features, we have formed a final data set for the task of detecting true and (or) false statements. The data set was divided into training, validation and test samples, taking into account the balance of classes and overall representativeness.

All the described materials and methods were implemented in order to create a reliable, realistic and informative data set for further research and development of models for detecting true and (or) false statements.

4 Results and discussion

1. Description of the data set. Our experimentally created data set for the task of detecting true and (or) false statements consists of 372 video fragments, which includes 170 false statements, 202 true statements.
2. Analysis of the results. We analyzed the data set to evaluate its characteristics and suitability for the task of detecting true and (or) false statements. During the analysis, we paid attention to the following aspects:
   - ○Class balance. We found that our dataset has a balanced ratio of true and false statements. This is important to ensure that the models are trained correctly and that reliable results are obtained.
   - ○Representativeness. The dataset covers a variety of scenarios and conditions that may occur in real life. This allows models to learn from a variety of examples and generalize to new data.
   - ○Markup quality. During the verification and validation of the data markup, we were convinced of the high accuracy and reliability of the truth labels of statements. This is essential for the correct evaluation of the performance of models and the accuracy of the results.

   Based on the preprocessed video materials, annotated data, and extracted features, we compiled a final dataset for the task of detecting true and (or) false statements. This dataset was carefully divided into training, validation, and test samples, ensuring a balance of classes and overall representativeness.
3. Discussion of results. The results of our experiments on the created data set confirm its value and suitability for the task of detecting true and (or) false statements. It can serve as an important tool for researchers engaged in the analysis of disinformation and the detection of fake news.

   However, it should be noted that our data set is not exhaustive and may be supplemented in the future. It is also worth considering that the results of our experiments may depend on the chosen algorithms and model parameters, and other approaches may lead to different results.

   In general, our data set and conducted experiments represent an important contribution to the field of automatic detection of true and (or) false statements and open up opportunities for further research and improvements in this area.

Acknowledgement

The research was conducted under the grant funding from the Ministry of Education and Science of the Republic of Kazakhstan as part of Project AP13068084 ”Development of technologies for detecting anomalous (deceptive) respondent behavior using artificial intelligence (AI) algorithms based on changes in voice and speech characteristics” (Competition for Young Scientists for Scientific and/or Scientific-Technical Projects for 2022-2024). We extend our heartfelt gratitude to the Ministry for their support and recognition of the significance of this research endeavor.

References

[1] Wu, Z., Singh, B., Davis, L., Subrahmanian, V. Deception Detection in Videos. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1)(2018). https://doi.org/10.1609/aaai.v32i1.11502

[2] Soldner, F., Pérez-Rosas, V., Mihalcea, R. Box of Lies: Multimodal Deception Detection in Dialogues. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 1768–1777, Minneapolis, Minnesota. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/N19-1175

[3] Burzo, M., Abouelenien, M., Pérez-Rosas, V., Mihalcea, R. Multimodal deception detection. The Handbook of Multimodal-Multisensor Interfaces, Volume 2, pages 419–453 (2018). https://doi.org/10.1145/3107990.3108005

[4] Gupta, V., Agarwal, M., Arora, M., Chakraborty, T., Singh, R., et al. Bag-of-Lies: A Multimodal Dataset for Deception Detection. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Long Beach, CA, USA, pages 83-90 (2019). https://doi.org/10.1109/CVPRW.2019.00016

[5] Su, L., Levine, M. High-Stakes Deception Detection Based on Facial Expressions. 2014 22nd International Conference on Pattern Recognition, Stockholm, Sweden, pages 2519-2524 (2014). https://doi.org/10.1109/ICPR.2014.435

[6] Lloyd, E., Deska, J., Hugenberg, K., McConnell, A., Humphrey, B., et al. Miami University deception detection database. Behavior Research Methods, 429-439 (2019). https://doi.org/10.3758/s13428-018-1061-4

[7] Radlak, K., Bozek, M., Smolka, B. Silesian Deception Database: Presentation and Analysis. WMDD ’15: Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection (WMDD ’15). Association for Computing Machinery, New York, NY, USA, pages 29–35 (2015). https://doi.org/10.1145/2823465.2823469

[8] Pérez-Rosas, V., Abouelenien, M., Mihalcea, R., Burzo, M. Deception Detection using Real-life Trial Data. Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, pages 59-66 (2015). https://doi.org/10.1145/2818346.2820758

[9] Mathur, L., Matarić, M. Unsupervised Audio-Visual Subspace Alignment for High-Stakes Deception Detection. ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, pages 2255-2259 (2021). https://doi.org/10.1109/ICASSP39728.2021.9413550

[10] Ji, S., Xu, W., Yang, M., Yu, K. 3D Convolutional Neural Networks for Human Action Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 1, pp. 221-231 (2013). https://doi.org/10.1109/TPAMI.2012.59

[11] Li, J., Wang, Y., Wang, C., Tai, Y., Qian, J., et al.DSFD: Dual Shot Face Detector. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, pages 5055-5064 (2019). https://doi.org/10.1109/CVPR.2019.00520

[12] Meng, D., Peng, X., Wang, K., Qiao, Y. Frame Attention Networks for Facial Expression Recognition in Videos. 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, pages 3866-3870 (2019). https://doi.org/10.1109/ICIP.2019.8803603

[13] Eyben, F., Weninger, F., Gross, F., Wolfgang Schüller, B. Recent developments in openSMILE, the munich open-source multimedia feature extractor. Proceedings of the 21st ACM international conference on Multimedia, pages 835–838 (2013). https://doi.org/10.1145/2502081.2502224

[14] Baltrusaitis, T., Zadeh, A., Lim, Y., Morency, L.-P. OpenFace 2.0: Facial Behavior Analysis Toolkit. IEEE International Conference on Automatic Face Gesture Recognition (FG), pages 59–66 (2018). https://doi.org/10.1109/FG.2018.00019

[15] Gogate, M., Adeel, A., Hussain, A. Deep learning driven multimodal fusion for automated deception detection. 2017 IEEE Symposium Series on Computational Intelligence (SSCI), pages 1-6 (2017). https://doi.org/10.1109/SSCI.2017.8285382

[16] Saito, M. Lie detection infiltrating everyday life. Retrieved from The Japan Times: https://www.japantimes.co.jp/life/2002/04/25/digital/lie-detection-infiltrating-everyday-life/ (2002).

[17] Taylan , S., Md Kamrul , H., Zach , T., Mohammed (Ehsan) , H. Automated Dyadic Data Recorder (ADDR) Framework and Analysis of Facial Cues in Deceptive Communication. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, pages 1-22 (2018). https://doi.org/10.48550/arXiv.1709.02414

[18] Venkatesh, S., Ramachandra, R., Bours, P. (2019). Robust Algorithm for Multimodal Deception Detection. 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), San Jose, CA, USA, pages 534-537. https://doi.org/10.1109/MIPR.2019.00108

[19] Yang, J.-T., Liu, G.-M., Huang, S.-H. Emotion Transformation Feature: Novel Feature For Deception Detection In Videos. 2020 IEEE International Conference on Image Processing (ICIP), pages 1726-1730 (2020). https://doi.org/10.1109/ICIP40778.2020.9190846

Aikumis Omirali received her Master's degree in Data Engineering from Narxoz University. She has a solid foundation in technical engineering, having obtained a Bachelor's degree in Computer Hardware and Software Engineering. Her professional responsibilities include data collection, handling day-to-day project tasks, and managing project documentation effectively. Her research interests focus on data analysis, data gathering from open sources, and the development and administration of databases. This combination of skills allows her to contribute substantially to the technical aspects of the projects she is involved in. She has published several papers that provide substantial backing for the research undertaken in her projects, particularly in innovative data engineering techniques that enhance project outcomes. In addition to her technical and research activities, Aikumis plays a crucial role in maintaining the informational infrastructure of the projects she works on, ensuring that the data collected is managed properly and effectively used to drive project success. Her work has been acknowledged in various scholarly publications, underscoring her contributions to the field of data engineering.

Rakhima Zhumaliyeva holds a PhD in software-based experimental-phonetic research with languages English, Kazakh, and Russian. She is a Research Philologist and has served as a Teacher of English and German. Her extensive experience encompasses both teaching and research activities. She has played a significant role in grant and program-targeted funding projects of the Ministry of Education and Science of the Republic of Kazakhstan (MES RK):

- AP05133600 "Developing and implementing an innovative competence-based model of multilingual IT specialist in the course of national education system modernization" (2018-2020);
- BR05236699 "Development of a digital adaptive educational environment using large-scale analysis of data" (2018-2020).

She is the author of 130 scientific works, including 3 monographs on the competency-based approach in IT-education and publications indexed in Scopus and Web of Science.

Didar Bayazitov received his Master's degree in Data Engineering from Narxoz University. He has a strong background in computer engineering and software, with a Bachelor's degree in Computer Hardware and Software Engineering. His research interests include data analysis, data scraping, and backend development, complemented by skills in frontend web development. This blend of expertise enables him to effectively manage data collection,

analysis, content uploading, and support for the project's website. He has published several articles that provide foundational research for this project, exploring innovative approaches in data engineering to enhance project outcomes. His work has been recognized in various technical publications, reflecting his contributions to the field of data engineering. In addition to his research activities, he plays an active role in supporting the informational infrastructure of project websites, ensuring that data-driven insights are accessible and impactful.

Kanat Kozhakhmet is the Project Leader and Chief Scientific Researcher appointed by the National Science Council in the field of "Information, Communication, and Space Technologies". He was the principal investigator for projects funded by the Ministry of Education and Science of Kazakhstan: BR05236699 "Development of a digital adaptive educational environment using large-scale data analytics" (2018-2020); AP05133600 "Development and implementation of an innovative competency model for polyglot IT specialists in the context of modernization of national education" (2018-2020). He has served as an expert at the National Center for Science and Technology Evaluation and is the chairman of the IT Alliance. He is a Visiting Scholar at prominent institutions, contributing to international research collaborations. He received his PhD in "Computer Science, Computing, and Management" from a leading institution. He is an editor and referee for several esteemed journals in the fields of artificial intelligence, machine learning, and information security. His main research interests are in artificial intelligence, neural language processing (NLP), machine learning, and information security. Additionally, he focuses on the implementation of innovative educational technologies and the development of professional standards and educational programs in IT.