Lecture Notes in Networks and Systems 1302

Abhishek Swaroop
Bal Virdee
Sérgio Duarte Correia
Zdzislaw Polkowski *Editors*

Proceedings of Data Analytics and Management

ICDAM 2024, Volume 6

Springer

Lecture Notes in Networks and Systems

Volume 1302

The series "Lecture Notes in Networks and Systems" publishes the latest developments in Networks and Systems—quickly, informally and with high quality. Original research reported in proceedings and post-proceedings represents the core of LNNS.

Volumes published in LNNS embrace all aspects and subfields of, as well as new challenges in, Networks and Systems.

The series contains proceedings and edited volumes in systems and networks, spanning the areas of Cyber-Physical Systems, Autonomous Systems, Sensor Networks, Control Systems, Energy Systems, Automotive Systems, Biological Systems, Vehicular Networking and Connected Vehicles, Aerospace Systems, Automation, Manufacturing, Smart Grids, Nonlinear Systems, Power Systems, Robotics, Social Systems, Economic Systems and other. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution and exposure which enable both a wide and rapid dissemination of research output.

The series covers the theory, applications, and perspectives on the state of the art and future developments relevant to systems and networks, decision making, control, complex processes and related areas, as embedded in the fields of interdisciplinary and applied sciences, engineering, computer science, physics, economics, social, and life sciences, as well as the paradigms and methodologies behind them.

Indexed by SCOPUS, EI Compendex, INSPEC, WTI Frankfurt eG, zbMATH, SCImago.

All books published in the series are submitted for consideration in Web of Science.

For proposals from Asia please contact Aninda Bose (aninda.bose@springer.com).

Abhishek Swaroop · Bal Virdee ·
Sérgio Duarte Correia · Zdzislaw Polkowski
Editors

Proceedings of Data Analytics and Management

ICDAM 2024, Volume 6

Springer

Editors
Abhishek Swaroop
Department of Information Technology
Bhagwan Parshuram Institute
of Technology
New Delhi, Delhi, India

Sérgio Duarte Correia
Polytechnic Institute of Portalegre
Portalegre, Portugal

Bal Virdee
Centre for Communications Technology
London Metropolitan University
London, UK

Zdzislaw Polkowski
Jan Wyzykowski University
Polkowice, Poland

ISSN 2367-3370 ISSN 2367-3389 (electronic)
Lecture Notes in Networks and Systems
ISBN 978-981-96-3380-7 ISBN 978-981-96-3381-4 (eBook)
https://doi.org/10.1007/978-981-96-3381-4

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Steering Committee

Patrons

Prof. Dr. Don MacRaild, Pro-Vice Chancellor, London Metropolitan University
Prof. Dr. Wioletta Boznańska, Rector, The Karkonosze State University of Applied Sciences in Jelenia Góra, Poland
Prof. Dr. Beata Telążka, Vice-Rector, The Karkonosze State University of Applied Sciences in Jelenia Góra, Poland

Honorary Chairs

Prof. Dr. Vaclav Snasel, Rector, VSB-Technical University of Ostrava, Czech Republic
Chris Lane, London Metropolitan University

General Chairs

Prof. Dr. Janusz Kacprzyk, Polish Academy of Sciences, Systems Research Institute, Poland
Prof. Dr. Karim Ouazzane, London Metropolitan University

Conference Chairs

Prof. Dr. Vassil Vassilev, London Metropolitan University
Dr. Pancham Shukla, Imperial College London

Prof. Dr. Mak Sharma, Birmingham City University
Prof. Dr. Anil K. Ahlawat, Dean, KIET Group of Institutes, India

Conveners

Prof. Dr. Bal Virdee, London Metropolitan University
Prof. Sérgio Duarte Correia, Politécnico de Portalegre, Portugal

Publication Chairs

Prof. Dr. Abhishek Swaroop, Professor, HoD, Bhagwan Parshuram Institute of Technology, India
Dr. Zdzislaw Polkowski, Department of Humanities and Social Sciences, The Karkonosze University of Applied Sciences in Jelenia Góra, Poland
Dr. Sandra Fernando, London Metropolitan University

Technical Program Chairs

Dr. Shahram Salekzamankhani, London Metropolitan University
Dr. Mohammad Hossein Amirhosseini, University of East London
Dr. Richa Sharma, London Metropolitan University
Dr. Qicheng Yu, London Metropolitan University
Dion Mariyanayagam, London Metropolitan University
Dr. Suresh Chandra Padhy, President, Poornima University, Jaipur, India
Dr. Jameel Ahamed, Maulana Azad National Urdu University, Gachibowli, Hyderabad, India
Dr. Moolchand Sharma, Maharaja Agrasen Institute of Technology (GGSIPU), New Delhi

Publicity Chairs

Dr. Józef Zaprucki, Prof. KPSW, Rector's Proxy for Foreign Affairs, The Karkonosze State University of Applied Sciences in Jelenia Góra
Dr. Umesh Gupta, Bennett University, India
Dr. Udit Mamodiya, Associate Dean (Research), Poornima University, Jaipur, India
Dr. Deepak Arora, Professor and Head (CSE), Amity University, Lucknow Campus
João Matos-Carvalho, Lusófona University, Portugal

Preface

We hereby are delighted to announce that London Metropolitan University, London, UK, in association with WSG University, Bydgoszcz Poland, Portalegre Polytechnic University, Portugal, Europe, and BPIT, GGSIPU, Delhi, has hosted the eagerly awaited and much coveted International Conference on Data Analytics and Management (ICDAM-2024) in Hybrid Mode on June 14–15, 2024. The fifth version of the conference was able to attract a diverse range of engineering practitioners, academicians, scholars, and industry delegates, with the reception of abstracts including more than 6000 authors from different parts of the world. The committee of professionals dedicated to the conference is striving to achieve a high-quality technical program with tracks on Data Analytics with Computer Networks. All the tracks chosen for the conference are interrelated and are very famous among the present-day research community. Therefore, a lot of research is happening in the above-mentioned tracks and their related sub-areas. As the name of the conference starts with the word 'Data Analytics', it has targeted out-of-the-box ideas, methodologies, applications, expositions, surveys, and presentations helping to upgrade the status of research. More than 1400 full-length papers have been received, among which the contributions are focused on theoretical, computer simulation-based research, and laboratory-scale experiments. Amongst these manuscripts, 300 papers have been included in the Springer proceedings after a thorough two-stage review and editing process. All the manuscripts submitted to the ICDAM-2024 were peer-reviewed by at least two independent reviewers, who were provided with a detailed review proforma. The comments from the reviewers were communicated to the authors, who incorporated the suggestions in their revised manuscripts. The recommendations from two reviewers were taken into consideration while selecting a manuscript for inclusion in the proceedings. The exhaustiveness of the review process is evident, given the large number of articles received addressing a wide range of research areas. The stringent review process ensured that each published manuscript met the rigorous academic and scientific standards. It is an exalting experience to finally see these elite contributions materialize into six book volumes as ICDAM-2024 proceedings by Springer entitled "International Conference on Data Analytics and Management". The articles are organized into six volumes in some broad categories covering subject matters on machine learning, data mining, data analytics, big data, networks, soft computing, and cloud computing, although given the diverse areas of research reported it might not have been always possible.

ICDAM-2024 invited five keynote speakers and eminent computer science and engineering researchers from around the world. In addition to the plenary sessions on each day of the conference, twenty-two concurrent technical sessions are held on both days to ensure the oral presentation of around 300 accepted papers. Keynote speakers and session chair(s) for each concurrent session have been leading researchers from the thematic area of the session. A technical exhibition is held during the 2 days of the conference, which displays the latest technologies, expositions, ideas, and presentations. The research part of the conference was organized in a total of 35 unique sessions. These special sessions and international workshops allowed researchers to conduct research in specific areas to present their results in a more focused environment.

An international conference of such magnitude and the release of the ICDAM-2024 proceedings by Springer have been the remarkable outcome of the untiring efforts of the entire organizing team. The success of an event undoubtedly involves the painstaking efforts of several contributors at different stages, dictated by their devotion and sincerity. Fortunately, since the beginning of its journey, ICDAM-2024 has received support and contributions from every corner. We thank all who have wished the best for ICDAM-2024 and contributed by any means towards its success. The edited proceedings volumes by Springer would not have been possible without the perseverance of all the steering, advisory, and technical program committee members.

All the contributing authors thank the organizers of ICDAM-2024 for their interest and exceptional articles. We would also like to thank the authors of the papers for adhering to the schedule and incorporating the review comments. We extend my heartfelt acknowledgement to the authors, peer-reviewers, committee members, and production staff whose diligent work shaped the ICDAM-2024 proceedings. We especially want to thank our dedicated peer reviewers who volunteered for the arduous and tedious step of quality checking and critiquing the submitted manuscripts. We thank my faculty colleagues, Dr. Moolchand Sharma, Dr. Jameel Ahmad and Dr. Simarpreet Singh, for extending their enormous assistance during the conference. The time spent by them, and the midnight oil burnt is greatly appreciated, for which we will ever remain indebted. The college's management, faculties, administrative, and support staff have constantly been extending their services whenever needed, for which we remain thankful to them.

We would like to express our sincere gratitude to Springer for accepting our proposal to publish the ICDAM-2024 conference proceedings. The support and guidance we received from Mr. Aninda Bose, the acquisition senior editor, were instrumental in making this a reality. We are truly grateful for their assistance and look forward to our continued collaboration.

Convener, ICDAM-2024

Contents

About the Editors

Prof. (Dr.) Abhishek Swaroop completed his B.Tech. (CSE) from GBP University of Agriculture and Technology, M.Tech. from Punjabi University Patiala, and Ph.D. from NIT Kurukshetra. He has industrial experience of eight years in organizations like Usha Rectifier Corporations and Envirotech Instruments Pvt. Limited. He has 22 years of teaching experience. He has served in reputed educational institutions such as Jaypee Institute of Information Technology, Noida, Sharda University Greater, Noida, and Galgotias University Greater, Noida. He has served at various administrative positions such as the head of the Department, division chair, NBA coordinator for the university, and head of training and placements. Currently, he is serving as a professor and HoD, Department of Information Technology in Bhagwan Parshuram Institute of Technology, Rohini, and Delhi. He is actively engaged in research. He has more than 60 quality publications, out of which eight are SCI and 16 Scopus.

Prof. Bal Virdee graduated with a B.Sc. (Engineering) Honors in Communication Engineering and M.Phil. from Leeds University, UK. He obtained his Ph.D. from the University of North London, UK. He worked as an academic at Open University and Leeds University. Prior to this, he was a research and development electronic engineer in the Future Products Department at Teledyne Defence (formerly Filtronic Components Ltd., Shipley, West Yorkshire) and at PYE TVT (Philips) in Cambridge. He has held numerous duties and responsibilities at the university, i.e., health and safety officer, postgraduate tutor, examination's officer, admission's tutor, short course organizer, course leader for M.Sc./M.Eng. Satellite Communications, B.Sc. Communications Systems, and B.Sc. Electronics. In 2010, he was appointed an academic leader (UG Recruitment). He is a member of ethical committee and member of the school's research committee and research degrees committee.

Prof. Sérgio Duarte Correia received his Diploma in Electrical and Computer Engineering from the University of Coimbra, Portugal, in 2000, the master's degree in Industrial Control and Maintenance Systems from Beira Interior University, Covilhã, Portugal, in 2010, and the Ph.D. in Electrical and Computer Engineering from the University of Coimbra, Portugal, in 2020. Currently, he is an associate professor at the Polytechnic Institute of Portalegre, Portugal. He is a researcher at COPELABS—Cognitive and People-centric Computing Research Center, Lusófona University of Humanities and Technologies, Lisbon, Portugal, and Valoriza—Research Center for Endogenous Resource Valorization, Polytechnic Institute of Portalegre, Portalegre, Portugal. Over past 20 years, he has worked with several private companies in the field of product development and industrial electronics. His current research interests are artificial intelligence, soft computing, signal processing, and embedded computing.

Prof. (Dr.) Zdzislaw Polkowski is an adjunct professor at Faculty of Technical Sciences at the Jan Wyzykowski University, Poland. He is also the rector's representative for International Cooperation and Erasmus Program and former dean of the Technical Sciences Faculty during the period of 2009–2012. His area of research includes management information systems, business informatics, IT in business and administration, IT security, small medium enterprises, CC, IoT, big data, business intelligence, and blockchain. He has published around 60 research articles. He has served the research community in the capacity of author, professor, reviewer, keynote speaker, and co-editor. He has attended several international conferences in the various parts of the world. He is also playing the role of principal investigator.

Detection of Pharmaceutical Pill Defects Through Deep One-Class Classification

Kunal Roy Choudhury, Animesh Singh, and S. Padmini

Abstract This paper proposes a deep learning classification method for detecting the defects in pills which is the day to day necessity of a pharmaceutical industry. This paper uses a Fully Convolutional Data Description (FCDD) network which detects the pill detects and provides an enhanced control on quality measures. Detection of anomaly in pills using FCDD plays a crucial role in the pharmaceutical industry by its visual inspection process and automatic quality check. The ability of the proposed algorithm proves its validation in detecting the pill image anomalies paving its way in meeting effective safety and stringent standards adhering to customer health and upholding industry regulations. Thus, this paper contributes to the path of achieving Sustainable Development Goal-3, respectively.

Keywords FCDD · Pharmaceutical manufacturing · Anomaly detection · ML · SDG

1 Introduction

1.1 Main Contribution

- FCDD involves detailed feature extraction process such as color variation, various in texture and other unique marking for analyzing images on pill images anomalies.

K. Roy Choudhury · A. Singh · S. Padmini (✉)
Department of Computing Technologies, SRM Institute of Science and Technology, Kattankulathur, Chennai, India
e-mail: padminis@srmist.edu.in

K. Roy Choudhury
e-mail: kr0296@srmist.edu.in

A. Singh
e-mail: as2957@srmist.edu.in

A. Swaroop et al. (eds.), *Proceedings of Data Analytics and Management*, Lecture Notes in Networks and Systems 1302, https://doi.org/10.1007/978-981-96-3381-4_1

- Integration of Multiple Data Types: FCDD integrates data from fluorescent labeling and colorimetric analysis, combining both types of information to enhance pill detection accuracy.
- FCDD uses Machine Learning (ML) algorithms for recognizing the different types of anomalies in the pill images based on its visual attributes.
- Optimized for real-time use, FCDD systems provide quick and accurate detection of pharmaceutical pills, crucial for manufacturing processes, quality control and identifying counterfeit drugs.
- FCDD captures miniature details and subtle differences in pill ages, enhancing high sensitivity and specificity for reliable identification.

2 Literature Survey

Paper [1, 2] employed Convolutional Neural Networks (CNNs) for accurate classification for defect in pills in the pharmaceutical industry. The authors in [3] employed real-time-based IoT monitoring system for effective pill defect detection. Paper [4] presented a unique technique for processing the images for anomaly detection in pills. Paper [5] proposed the AI models capable of defective pill detection. Paper [6, 7] employed the technique of computer vision for analyzing tablet disintegration. Paper [8] presented a combined approach of neural networks and computer vision for real-time defect in pill identification. Paper [9] covered non-destructive testing techniques for assessing pharmaceutical tablet quality, focusing on defect identification methods. Wang and Gupta [10] investigated hyper spectral imaging technology for pharmaceutical quality control, enhancing defect detection capabilities through advanced imaging. Using sensor data, a study [11] focused on anomaly detection in tablet manufacturing, contributing to early defect detection and improved manufacturing quality. Paper [12, 13] presents the viabilities and challenges available in the automatic pill detection in the pharmaceutical industry, spotlighting the need for hi-tech technologies in the detection of pill defect processes. Accuracy improvement in classifying the defective pills has been effectively analyzed by the use of transfer learning technique [14]. An inclusive survey [15] listed the practices in pharmaceutical quality assurance practices complying to industry regulatory measures.

3 Proposed Methodology

The FCDD algorithm utilizes a PillQC dataset. This algorithm employs the probability score of the heat map to classify the defect in the images of the pill. The dataset, with 149 normal, 43 chip-defected and 138 dirt-contaminated images, prepares the model for diverse pill anomalies. Preprocessed and categorized images facilitate training, with the system addressing imaging issues like shadows and blurring. It

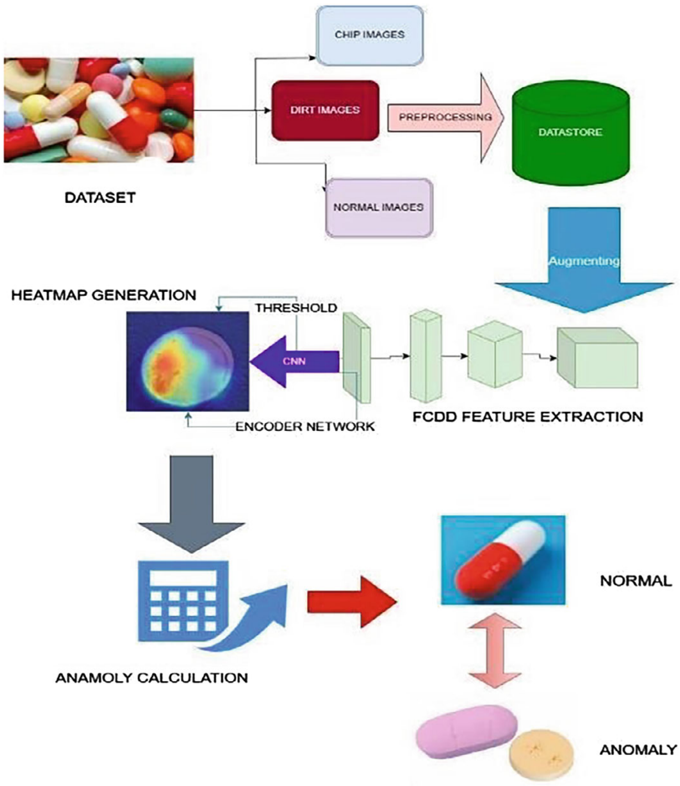

Fig. 1 Overall system architecture

emphasizes anomaly areas for improved interpretability and detection accuracy. Figure 1 shows the overall architecture of the system.

3.1 Data Preprocessing and Augmentation

For the FCDD model, images are loaded into an image Datastore and automatically labeled, then split into training, calibration, and test sets to ensure model accuracy. Data augmentation techniques like rotations and flips mimic real-world conditions. The model is trained primarily on normal images, with the calibration set fine-tuning the anomaly threshold and the test set verifying defect detection. Custom functions like 'split Anomaly Data', 'augment DataForPillAnomaly Detector' and 'addLabelData' aid in organizing and enhancing the dataset for effective training and testing.

3.2 Using FCDD Data Model

The Fully Convolutional Data Description (FCDD) uses Fully Convolutional Networks (FCNs) and the Hypersphere Classifier (HSC) for deep one-class anomaly detection. It maintains spatial details and creates a reduced-resolution anomaly heatmap. The FCDD analyzes the average heatmap value to classify pixels and overall images as normal or abnormal.

The heatmap A(X) in the FCDD model is created using strided convolution with a Gaussian kernel, whose mean and variance depend on the input pixel at the center of the receptive field for a given output pixel. Although the specific up-sampling formula using the Gaussian kernel is not detailed, the procedure includes convolving A(X) with this kernel, aiming to amplify anomaly scores for actual anomalies and reduce them for normal instances. For Convolutional Neural Networks (CNNs), the dimension of the output feature map, given an input image of dimension of n*n and a filter size $f*f$, is generally computed by the formula that incorporates these dimensions along with stride and padding factors given by:

$$((\mathrm{n}-\mathrm{f}+1),(\mathrm{n}-\mathrm{f}+1)) \tag{1}$$

The FCDD model merges Fully Convolutional Network (FCN)-based feature extraction with an innovative adaptation of the HSC loss for anomaly detection, creating heatmaps to pinpoint anomalies in images. It operates by training to produce an anomaly score map, assessing each image region for potential anomalies. The model's foundation uses the first three down- sampling stages of a pre-trained Inception-v3 network from ImageNet, forming the backbone of the FCDD.

$$\begin{aligned}&\text{Accuracy} = (\text{Number of correct predictions})/(\text{Total number of predictions})\\&\text{Precision} = (\text{True positives })/(\text{False positives} + \text{ truepositives})\end{aligned} \tag{2}$$

FCDD encompassed unique procedure unlike traditional methods, that excel in spatially extraction of features . This informs intricate details within pill images such as color, texture and imprint distributions. Thus, this method leads to more robust and accurate characterization of pill images. Further its ability to handle variable input sizes, where pills may vary in dimensions are crucial in pharmaceutical industries. This flexibility adhered with its hierarchical architecture enables with multi-level feature representation, provides the system to learn and understand complex patterns and enables informed detections. Furthermore, FCDD can be utilized for real-time monitoring, providing rapid feedback in assuring the quality control processes. Its inherent integration with multiple modalities enhances its capability to achieve accuracy of higher detection. FCDD's suitability to diverse pill types, high sensitivity to miniature variations and automated decision-making capabilities further highlights its unique robustness in defective pill detection pharmaceutical industry. The proposed workflow of the FCDD model is shown in Fig. 2.

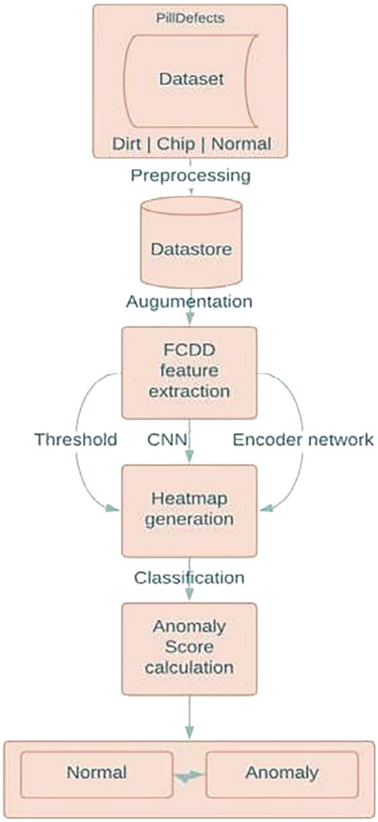

Fig. 2 Proposed FCDD model workflow

4 Simulation Result

PillQC dataset is taken as the test set for performing the anomaly detection. Here a confusion matrix offers solution for visualization of performances as normal or abnormal classification. This methods helps to find the accuracy as well as errors in classification. Matrix method using functions has been adopted to find the classification accuracy of the proposed system. Table 1 gives the accuracy per sub-class for the different types of experimental data taken from the dataset.

Table 2 gives the proposed method outperforming R-CNN and YOLOV3 in accuracy. The model has higher accuracy for 'Normal' and slightly lesser for 'Anomaly,' with detailed accuracies for sub-classes 'chip' and 'dirt.' These findings, based on a

Table 1 Accuracy for sub-class data

Type	Accuracy per sub-class
Chip	0.84375
Dirt	0.99029

Table 2 Validation of the proposed method

Method	Accuracy	Precision
R-CNN [15]	0.916	0.916
YOLOV3 [15]	0.733	0.925
Proposed method	0.96923	0.985

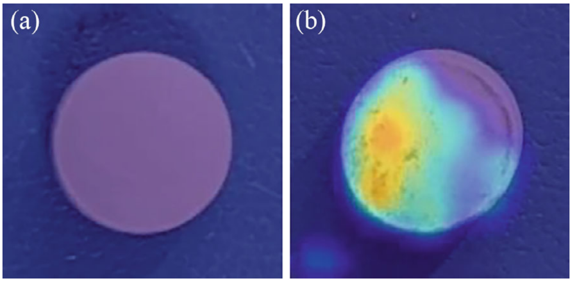

Fig. 3 **a** Pill image heat map3. **b** Pill image

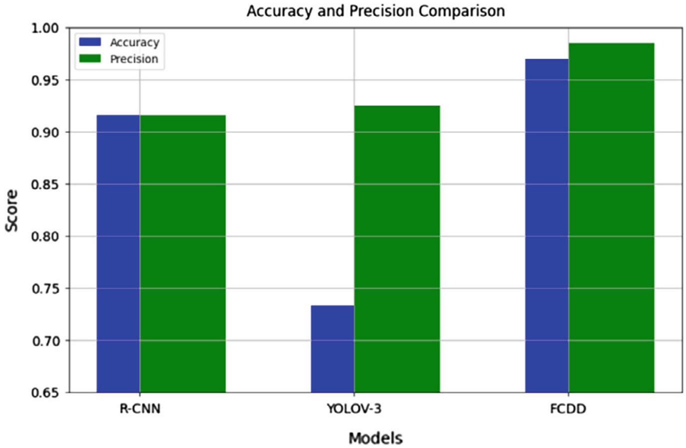

Fig. 4 Comparison of different models

single pill image, are also visually supported by a heatmap in Figs. 3 and 4 proposed the validation of the proposed model with the existing state of art [15].

5 Conclusion

The methodology of the project effectively handles image acquisition issues like shadows, focus blurring and color variation, showcasing the FCDD network's ability to identify anomalies in pill images, with potential use in medical diagnostics. The project highlights the strengths of one-class learning in situations where one-class dominates, contributing significantly to image anomaly detection technology. The model's performance is outstanding, with its ability to separate normal and anomalous instances, proven by a histogram of anomaly scores and optimized through the 'anomaly threshold' function. The above results interprets the model excels in distinguishing between normal and anomalous instances, promising broader applicability and effectiveness in anomaly detection. FCDD offers practical benefits in pharmaceutical quality assurance by enhancing visual inspection, automating processes, and contributing to achieving SDG-9 objectives. By leveraging advanced neural network architecture and automation capabilities, FCDD improves efficiency, reduces errors and ensures the consistent quality of pharmaceutical products, ultimately advancing the sustainability and resilience of the pharmaceutical industry.

References

1. Kim D, Park H (2019) Deep one-class pill detection with auto encoder pretraining. J Healthcare Eng1–9
2. Pang Y, Huang W, Chen Y (2020) A novel one-class method in anomaly detection on imbalanced data. IEEE Access 8:26391–26398
3. Moazami M, Tizhoosh HR (2020) Pill detection in medical images using one-class extreme learning machines. J Imaging 6(6):65
4. Gao Y, An B (2019) A survey of deep one-class classification. Int J Automat Comput 16(5):532–544
5. Yang J, Wu Z (2019) Semi-supervised anomaly detection for pill recognition in medical images. J Biomed Inform 98:103284
6. Al-Antari MA, Al-Masni MA, Mohammed M (2016) Pill identification using one-class classifiers ensemble. IEEE Access 4:9931–9943
7. Bergmann P, Meinke A, Rätsch G (2020) Single class deep novelty detection with a deep conditional generative model. *arXiv preprint*arXiv:2004.12945
8. Tahmasbi A, Ghassemian H (2020) Pill detection in medical images using one class convolutional restricted boltzmann machines. Pattern Recogn Lett 133:151–157
9. Shirazi SH, Charkari NM (2017) An ensemble of one-class classifiers for pill detection in x-ray images. J Med Syst 41(11):80
10. Sato Y, Kimura M (2015) Online novelty detection on temporal data. Expert Syst Appl 42(4):1849–1857
11. Razzak MI, Naz S (2020) A review of pill detection approaches using one-class classification. J Imag 6(5):52
12. Xu S, Chen B (2020) Feature learning in one-class classification problem. Neuro Comput 155:110–119
13. Fathi A, Faradji F, Neamati N (2019) A comparative study on one-class machine learning techniques for pill recognition. J Med Imag Health Inform 9(8):1640–1646
14. Zhou Z, Yhang Y, Sun H (2017) A deep based method for anomaly detectionin vibration signals. Sensors 17(11):2554
15. Kwon HJ, Kim HG, Lee SH (2022) Pill detection model for medicine inspection based on deep learning. J Conferen

Advanced Lip-Reading Techniques: Leveraging ST-CNN-BiGRU and CTC-Greedy Decoding for Improved Text Transcription

A. Kavya Varshini, C. Jayaashri, and S. Padmini

Abstract This study introduces an advanced lip-reading technology that transforms video streams into textual content. It utilizes Spatio-Temporal Convolutional Neural Networks (ST-CNNs), Bidirectional-Recurrent Neural Networks (Bi-RNNs), and Connectionist Temporal Classification (CTC) to effectively calculate loss and capture the dynamics of visual speech. The ST-CNN processes both spatial and temporal aspects, while Bi-RNN layers interpret sequence dependencies. The integration of CTC loss helps optimize the training process by aligning video frames with corresponding text efficiently. This trainable system is highly adaptable and accurate, managing different speech patterns and environmental conditions with ease. Tests show significant improvements over traditional LSTM-based models. The model reports a Character Error Rate (CER) of 2.3% and Word Error Rate (WER) of 5.0%, showing substantial enhancements from the previous 11.6% WER with 2D models. These results confirm the model's advanced capabilities in recognizing visual speech. Extensive evaluations confirm its effectiveness, providing a robust solution for real-time video-to-text conversion. Furthermore, the development of this technology supports the goals of SDG 10, that aims to improve socio-economic and political inclusion for all, including those with hearing disabilities, by reducing communication barriers.

Keywords Deep learning · Sequence-to-sequence models · Lip reading · Video processing · Transformer networks · Attention mechanism · Visual speech recognition

A. Kavya Varshini · C. Jayaashri · S. Padmini (✉)
Department of Computing Technologies, SRM Institute of Science and Technology, Kattankulathur, Chennai, India
e-mail: padminis@srmist.edu.in

A. Kavya Varshini
e-mail: kv0380@srmist.edu.in

C. Jayaashri
e-mail: cj7032@srmist.edu.in

A. Swaroop et al. (eds.), *Proceedings of Data Analytics and Management*, Lecture Notes in Networks and Systems 1302, https://doi.org/10.1007/978-981-96-3381-4_2

1 Introduction

Machine learning advances with a deep neural network model for lip reading, enhancing communication between visual and verbal mediums. This model processes video frames with ST-CNN layers, transforming spatiotemporal lip movement features into numerical representations. These are fed into a Bi-RNN, generating vectors indicating prediction probabilities. CTC, along with GIS, produces a text prediction.

The applications are transformative, particularly in non-verbal communication. In noisy or silent scenarios, this model enables clear communication and improves digital content accessibility for individuals with hearing impairments. Additionally, it supports real-time speech-to-text conversion, revolutionizing information access and consumption.

Figure 1 provides an overview of a complete lip-reading system. It begins with 'Input Video Frames,' where video sequences are fed into the system. These data are then fed to ST-CNN so that temporal and spatial movements are captured. Movement of the lip sequence is then captured by the RNN layer. CTC loss is then evaluated to represent the text as output. The final results are then compared with benchmarking results to attain accuracy.

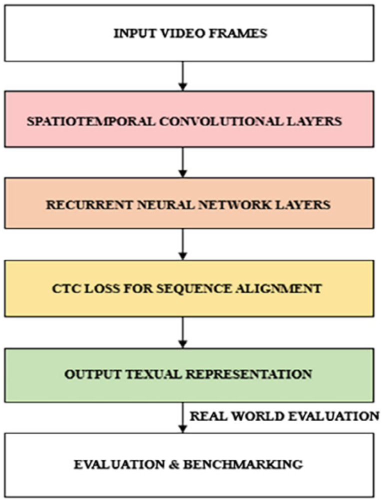

Fig. 1 Model evaluation process

2 Literature Survey

In paper [1, 2], different strategies and theory of chaos integration are embedded for accurate movement of lip conversion as text. Paper [3] explains the need of Recurrent Neural Networks (RNNs) for spatial and temporal sequence capturing of lip movements. Paper [4–6] presents the application of neural network in the current state of art. Paper [7] explores the usage of lip segmentation using Temporal-Convolutional Networks (TCNs). The combined usage of LSTM and RNN [8] for lip movement textual conversion has been represented in paper [8]. The LSTM blended with Convolutional Auto Encoders (CAE) for the futuristic approach has been proposed in paper [9].

Recent research in [10] introduces an attention-based LSTM and CNN lip-reading system with high precision, while [11] addresses keyword recognition challenges. Paper [12] focuses on enhancing recognition for new speakers and [13] explores radar sensing data for lip-reading, indicating a move toward alternative data sources. These studies highlight the evolving landscape of lip-reading and automatic speech recognition algorithms, aiming for more precise, robust, and efficient solutions..

3 Proposed Methodology

The dataset utilized for this project was gathered from the GRID audio-visual sentence corpus, a dataset that is well-known in the areas of speech recognition, lip reading, and research into audio-visual speech processing (Fig. 2).

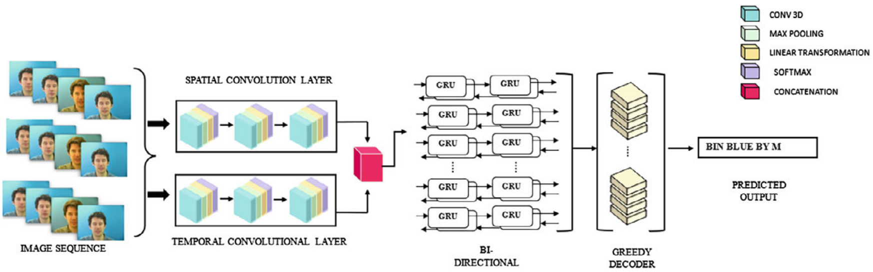

Fig. 2 Illustration of the proposed workflow

3.1 Spatio-Temporal Convolutional Neural Networks

Processing Inputs through ST-CNNs:

1. Preprocessing video data involves resizing frames, normalizing pixel values, and optionally converting them to grayscale. The clips are fed into an ST-CNN, which uses 3D convolutional layers to analyze spatial and temporal dimensions, capturing patterns and changes over time.
2. Spatial convolutions pinpoint crucial lip movement features for speech recognition, while temporal convolutions track lip movement dynamics for identifying speech rhythms and word distinctions.
3. As data moves through the network, ST-CNN layers extract increasingly complex features, from simple movements to detailed lip movement patterns. Pooling layers reduce feature map dimensionality, boosting computational efficiency, and model robustness.
4. The final ST-CNN layer produces a high-level representation of lip movement features, used in subsequent modeling and prediction stages, such as with LSTM networks. Given a 3D input tensor X of shape (L, P, Q), , where L represents temporal depth (number of frames), P represents height, and Q represents the width of the input, a 3D convolution operation involves a kernel or filter K of shape $\left(d_l, d_{p,} d_q\right)$, where d_l, d_p, and d_q correspond to the dimensions of the kernel in the temporal, height, and width dimensions, respectively.

The output of the 3D convolution operation at a specific location (l, p, q) in the output tensor Y can be mathematically expressed as:

$$Y(l, m, n) = \sum_{s=0}^{d_l-1} \sum_{t=0}^{d_p-1} \sum_{u=0}^{d_u-1} X(L + s, P + t, Q + u) \cdot K(s, t, u) \tag{1}$$

Using ST-CNNs, the model leverages spatial and temporal video data for accurate sentence-level end-to-end lip reading. Unlike traditional methods using handcrafted features or separate models for spatial and temporal analysis, this approach offers a cohesive and effective solution for interpreting speech from visual information alone.

3.2 Recurrent Neural Network with LSTM

Processing Inputs through RNNs with LSTM:

1. After processing video frames, the ST-CNN generates a high-level representation of the spatial and temporal features of lip movements, structured as a sequence with each element representing features from consecutive frames.
2. This feature sequence is input into an LSTM-based RNN, where each LSTM unit processes an element at a time, maintaining and updating an internal state that

summarizes previously seen information, facilitating the progression of relevant data through the sequence.

3. As the LSTM processes the sequence, it produces outputs at each step, which can predict the current video segment's content (e.g., likely phonemes or words from observed lip movements). These outputs are used for the final speech recognition task.

The output at each time stept, *y_t*, is then computed using both the forward and backward hidden states:

$$y_t = g\left(V\left[\overrightarrow{h_t}\,;\,\overleftarrow{h_t}\right] + c\right) \tag{2}$$

- V is the weight matrix for hidden-to-output connections.
- $\left[\overleftarrow{h_t}\,;\overrightarrow{h_t}\right]$ denotes the concatenation of the hidden states of backward and forward pass.
- c is the output bias.
- g is the output activation function, which can vary depending on the task (e.g., softmax for classification).

4. The transcription of the final text is enabled by the mapping between LSTM and CTC layer.
5. The interpretations of the movements of the lip are facilitated by the combined integration of LSTM with RNN model.

3.3 Connectionist Temporal Classification Layer

Processing Inputs through CTC:

1. The output from LSTM layer is fed to the CTC layer.
2. During training, the CTC layer computes CTC loss by text sequence based on LSTM outputs by computing its probability, considering all possible alignments between the predicted and target sequences.
3. The total probability of a target sequence y given the input X is the sum of the probabilities of all possible alignments π that can be collapsed into z (after removing blanks and merging repeated labels):

$$P(z|X) = \sum_{\pi \in \phi(z)} P(\pi|X) \tag{3}$$

where $\phi(z)$ is the set of all possible mappings that correspond to z.

CTC loss is an indicator of how closely the model's predictions match the target text guiding the model's training process to improve its predictions. The CTC loss for the frequency y given the input sequence X

$$\mathcal{L}_{CTC}(X, y) = -\log P(y|X) \tag{4}$$

This loss is minimized during training to improve the alignment between the input sequences and their corresponding target sequences.

4. During inference, the CTC layer decodes the LSTM output probabilities to produce the most likely text sequence, using strategies like greedy decoding.

3.4 Data Enhancement

The output data is enhancing after normalization and standarization of input data.

4 Simulation Results

The following performance metrics used for the current state of art is shown below by the calculation of Word Error Rate (WER) and Character Error Rate(CER):

$$WER = \frac{E + F + G}{H} \tag{5}$$

$$CER = \frac{E + F + G}{h} \tag{6}$$

Here, 'E' and 'e' indicate the model correction substitution for a word. Accordingly, 'F' and 'f' are the error removers in a word carried out by deletion. 'G' and 'g' are the error correction insertions for a word. 'H' and 'h' indicate the number of words in the transcript.

Table 1 use WER and CER as metrics to compare various speech recognition models on their performance. The Foundational LSTM, Foundational 2D model, and the Foundational NoLM model show a WER of 26.3%, 11.6% and 5.6% and the Foudational NoLm model has a CER of 2.4%. It could be observed that the Foundational NoLM model has significantly reduced the WER. The proposed model shows a WER of 4.8% and CER of 2.3%, it proves that the proposed model has a significantly less error rate than the other models it has been compared. This shows the models ability to accurately recognize character and words.

Table 1 Performance metrics

Error metrics	Foundational LSTM [14]	Foundational 2D [14]	Foundational NOLM [14]	Proposed model
WER	26.3%	11.6%	5.6%	4.8%
CER	–	–	2.4%	2.3%