ELYAN, E., MORENO-GARCÍA, C.F. and JOHNSTON, P. 2020. Symbols in engineering drawings (SiED): an imbalanced dataset benchmarked by convolutional neural networks. In Iliadis, L., Angelov, P.P., Jayne, C. and Pimenidis, E. (eds.) *Proceedings of the 21st Engineering applications of neural networks conference 2020 (EANN 2020); proceedings of the EANN 2020, 5-7 June 2020, Halkidiki, Greece.* Proceedings of the International Neural Networks Society, 2. Cham: Springer [online], pages 215-224. Available from: https://doi.org/10.1007/978-3-030-48791-1_16

Symbols in engineering drawings (SiED): an imbalanced dataset benchmarked by convolutional neural networks.

ELYAN, E., MORENO-GARCÍA, C.F. and JOHNSTON, P.

2020

This is an Author Accepted Manuscript version of the following chapter: ELYAN, E., MORENO-GARCÍA, C.F. and JOHNSTON, P. 2020. Symbols in engineering drawings (SiED): an imbalanced dataset benchmarked by convolutional neural networks. In Iliadis, L., Angelov, P.P., Jayne, C. and Pimenidis, E. (eds.) Proceedings of the 21st Engineering applications of neural networks conference 2020 (EANN 2020); proceedings of the EANN 2020, 5-7 June 2020, Halkidiki, Greece. Proceedings of the International Neural Networks Society, 2. Cham: Springer [online], pages 215-224. Available from: https://doi.org/10.1007/978-3-030-48791-1_16

This pre-copyedited version is made available under the Springer terms of reuse for AAMs: https://www.springer.com/gp/open-access/publication-policies/aam-terms-of-use

OpenAIR @RGU

This document was downloaded from https://openair.rgu.ac.uk

SEE TERMS OF USE IN BOX ABOVE

DISTRIBUTED UNDER LICENCE

Symbols in Engineering Drawings (SiED): An Imbalanced Dataset Benchmarked by Convolutional Neural Networks

Eyad Elyan[1], Carlos Francisco Moreno-García[1], and Pamela Johnston

The Robert Gordon University, Garthdee Road, Aberdeen, UK
e.elyan@rgu.ac.uk, c.moreno-garcia@rgu.ac.uk,p.johnston2@rgu.ac.uk

Abstract. Engineering drawings are common across different domains such as Oil & Gas, construction, mechanical and other domains. Automatic processing and analysis of these drawings is a challenging task. This is partly due to the complexity of these documents and also due to the lack of dataset availability in the public domain that can help push the research in this area. In this paper, we present a multiclass imbalanced dataset for the research community made of 2432 instances of engineering symbols. These symbols were extracted from a collection of complex engineering drawings known as Piping and Instrumentation Diagram (P&ID). By providing such dataset to the research community, we anticipate that this will help attract more attention to an important, yet overlooked industrial problem, and will also advance the research in such important and timely topics. We discuss the datasets characteristics in details, and we also show how Convolutional Neural Networks (CNNs) perform on such extremely imbalanced datasets. Finally, conclusions and future directions are discussed.

Keywords: CNN, Multiclass, Classification, Imbalanced Dataset, Engineering Drawings, P&ID

1 Introduction

Engineering drawings are known to be one of the most complex types of documents to process and analyse. They are widely used in different industries such as construction and city planning (i.e. floor plan diagrams [2]), Oil & Gas (i.e. P&IDs [9]), Mechanical Engineering [33], AutoCAD Drawing Exchange Format (DXF) [13] and others. Interpreting these drawings requires highly skilled people, and in some cases long hours of work. Processing and analysing these drawings is becoming increasingly important. This is partly due to the urgent need to improve business practices such as inventory, asset management, risk analysis, safety checks and other types of applications, and also due to the recent advancements in the domain of machine vision and image understanding. Deep Learning (DL) [15], in particular, had significantly improved the performance by orders of

magnitude in many domains such as the Gaming and AI [17], Natural Language Processing [36], Health [12], Cyber Security [28], and others.

The concept of Convolutional Neural Networks (CNNs) [16] has made significant progress in recent years in many image-related tasks. It has been successfully applied to several fields such as hand-written digit recognition [22], image classification [30,20], face recognition & biometrics [27], amongst others. Before CNNs, improvements in image classification, segmentation, and object detection were marginal and incremental. CNNs revolutionised this field. For example, Deep Face [31], which is a face recognition system that was first proposed by Facebook in 2014, achieved an accuracy of 97.35%, beating the then state-of-the-art, by 27%.

Despite extensive progress in the field of image processing and analysis, very little progress has been made in the area of analysing complex engineering drawings, and extracting information from these diagrams is still considered a challenging problem [5]. Consider for example the case of the Piping and Instrumentation Diagram (P&ID), which is a schematic engineering drawing, commonly used in the Oil and Gas industry [9,24]. This type of diagram, as can be seen in Figure 1, is made of symbols, connectivity information (lines, dashed lines, combinations of lines), text, and other graphical elements.

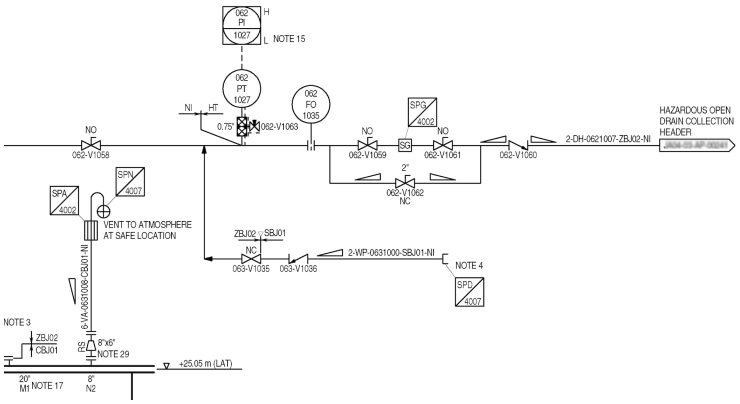

Fig. 1: A typical example of elements within a P&ID diagram

Identification of the symbols within this kind of diagram would appear to be an ideal problem which could be easily solved by convolutional neural networks. However, a recent review on the subject [7] showed that publicly available datasets are not common in this area, with research commonly applied to small, proprietary datasets. To take full advantage of the recent advances in machine vision, and to facilitate reproducible experiments, a sizeable, labelled dataset in the public domain is required.

Several factors make processing and analysing engineering drawings a challenging tasks. First, the quality of the images/scanned documents is sometimes of a standard which requires the application of various image-enhancements methods. Second, the nature of these diagrams, where various types of elements might be overlapping (i.e. a text overlaid on a symbol), in addition to possible data annotations and other graphic elements makes accurate localisation of individual elements more challenging. It is difficult to isolate one particular symbol from its neighbours. Another inherent problem is the imbalanced distribution of various symbols within these drawings. Handling all related challenges is beyond the scope of this paper. The reader is referred to [24] for more detailed description about the inherent characteristics and challenges of these types of drawings.

In this paper, we present a new multiclass dataset of symbols extracted from engineering drawings to the research community. Realistically reflecting the problem, this dataset is subject to some class-imbalance. The remaining parts of this paper are organised as follows: In Section 2 we discuss relevant literature to the digitisation of engineering drawings and class imbalance. In Section 3 we present our methods which includes detailed discussion of the dataset, and our approach for classifying engineering symbols. Benchmarking experiments and results are presented in Section 4, and finally, conclusions and future directions are discussed in Section 5.

2 Related Work

Attempts to process and analyse symbolic drawings date back to at least the early 90's. These include: analysis of musical notes [6]; processing mechanical drawings [19]; and optical character recognition (OCR)[21,23,26]. In recent years, digitising engineering drawings has become increasingly important as they are widely used in different domains [9,2,33,13], however, literature is still limited. To the best of our knowledge, there is no large, publicly available dataset to facilitate the advantages of modern, data-hungry CNNs. A recent review [7] detailed the whole process of digitisation and contextualisation of the three main shapes contained in engineering drawings (i.e. text, lines and symbols). The authors identified that, typically, symbols are located within the drawing either in a *specific* or a *holistic* way. In *specific* localisation, the system has a predefined symbol description/template, and an algorithm recursively looks for such symbol. In contrast, *holistic* methods require differentiation of the three shapes to then be able to split the drawing into layers. One of the most widely-used frameworks in this regard is text-graphics separation [32], which is a family of algorithms which distinguish text from lines and symbols based on properties such as height-to-width radio, stroke, amongst others. CNNs could be applied to both of these, given sufficient labelled data.

One type of engineering drawings, namely P&IDs, has attracted more research attention in recent years. Typical examples, presented in [9,18,25], aimed at detecting and recognising symbols within these diagrams. It can be argued however, that most of the existing literature followed a traditional image process-

ing approach [14], which requires feature extraction [8], feature representation [37], and classification to determine the class of objects (i.e. symbols, digits, ...) [1].

Most recently in [9], authors presented a first step towards creating a symbol repository for engineering drawings. A total of 1187 symbols split into 37 different classes was compiled. The repository was then processed by means of class decomposition [10,11], resulting in a total of 57 sub-classes. Classification accuracy was calculated using three different classification frameworks: Random Forests (RF), Support Vector Machine (SVM) and a CNN. Class decomposition demonstrated a slight improvement in classification results for SVM and CNN, with a more considerable improvement in RF.

Overall, there is a growing interest in the research community in digitising and analysing engineering drawings. Yet, the lack of public domain datasets is considered as one of the main challenges to push the research boundaries in this area. In addition to this, the class-imbalance problem could also be considered as another challenge, in particular when certain types of symbols either dominate, or rarely appear in the dataset. Class-imbalance is common across different domains, and not only limited to engineering drawings [35,34]. Handling this problem is often done by means of data resampling, where majority classes are undersampled to reduce their dominance, or minority classes are oversampled [35]. In addition to this, Generative Adversarial Neural Networks [15] were successfully applied recently to augment an imbalanced dataset and improve learning algorithm performance [3,4].

3 Methodology

This section presents our novel dataset of Symbols in Engineering Drawings (SiED). First we give a brief description of how this dataset was constructed. This is followed by a detailed description of dataset and class distribution. Finally, a brief discussion related to the classification method used to benchmark this dataset is presented.

3.1 Data Extraction

A collection of P&ID sheets was provided by an industrial partner. Following the work in [25], a thresholding method was first applied to reduce noise. Areas of interest were then identified interactively to discard boundaries, text and annotation outside the border of each drawing. A traditional machine-vision approach was then used to extract a set of symbols. A set of heuristic-based methods were developed and applied sequentially to localise symbols within each P&ID drawing. Figure 2 shows a random selection of typical symbols that appear in P&IDs.

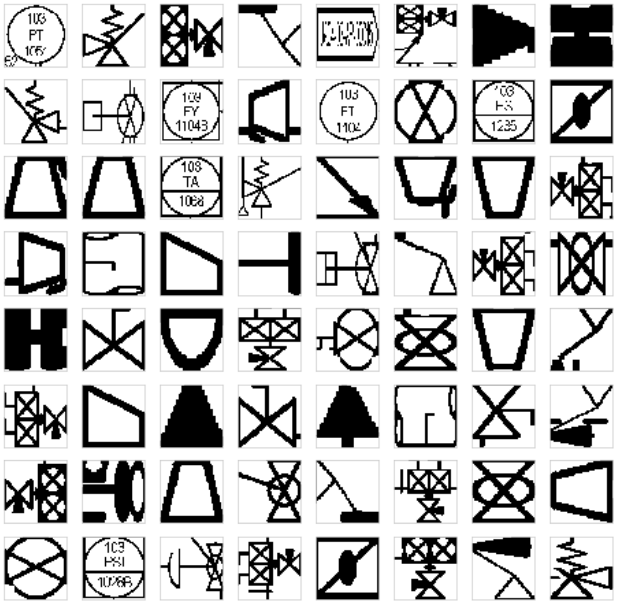

Fig. 2: A random selection of typical symbols that appear in P&IDs

The methods proved to be stable enough to provide a list of extracted and well-defined symbols. However, a key limitation of such heuristic-based methods is that they require extensive feature engineering and require fine-tuning and customisation to generalise to unseen symbols or different types of diagrams [7].

3.2 Dataset

Using the method presented above, a series of P&IDs have been processed and analysed. This resulted in a collection of symbols that represent different types of equipment within the drawings. In total, a dataset of 2432 instances representing 39 different type of symbols were compiled. All symbols have been scaled to a standard size of 100×100 pixels. The dataset provides rich source of information to evaluate various supervised machine learning algorithms. However, and as can be seen in Figure 3, the dataset is hugely imbalanced. Some symbols, such as sensors, dominate the dataset, while others appear only once or are vastly underrepresented.

The imbalance between symbols is huge in some cases. For example symbols of type *sensor* appears 392 times in the dataset, while symbols such as *Barred Tee* and *Ultrasonic Flow Meter* appear only once. Similarly, *Reducer* appears in the dataset 285 times, while *Control Valve Angle Choke* only once.

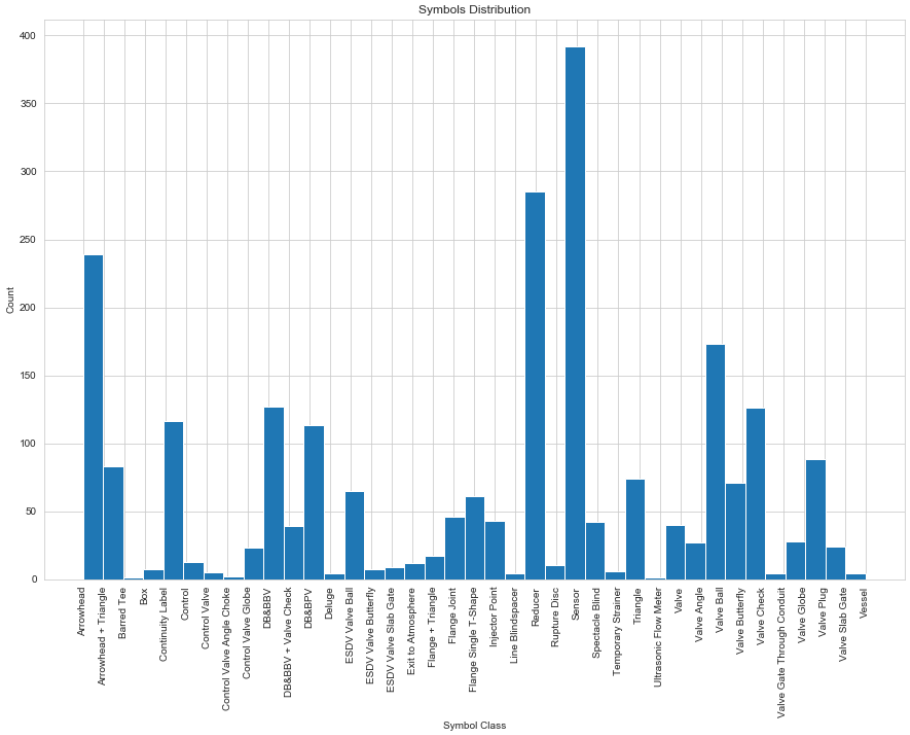

Fig. 3: Class distribution in the dataset

Interestingly, eight types of symbols populate more than 64% of the dataset. These are: Sensor, Reducer, Arrowhead, Valve Ball, DB&BBV, Valve Check, Continuity Label and DB&BPV. At the same time, 18 symbols populate together less than 6% of the whole dataset. These are: Valve Slab Gate, Control Valve Globe, Flange + Triangle, Control, Exit to Atmosphere, Rupture Disc, ESDV Valve Slab Gate, Box, ESDV Valve Butterfly, Temporary Strainer, Control Valve, Valve Gate Through Conduit, Deluge, Vessel, Line Blindspacer, Control Valve Angle Choke, Barred Tee and Ultrasonic Flow Meter. In other words, the dataset is hugely imbalanced.

3.3 Classification Method

To provide base-line results on this imbalanced dataset of symbols, we use CNNs. CNNs [16] have made significant progress in recent years in many image-related tasks and in particular in image classification [30,20].

The network architecture used in this paper consists of an input layer of 100×100 of the raw pixel values of the symbol and 32 filters (3×3). Then a 2×2 max pooling layer. Then, two convolutional layers followed by a 2×2 max pooling layer. This structure is then repeated twice with two convolutional

layers, with 64 filters of size (3×3) followed by a max pooling layer. Finally, a fully-connected layer composed of two hidden layers and an output layer of 39 (number of classes) units with softmax activation function. All convolutional layers in the network used *ReLU* activations. Dropout [29] was used in the in the fully connected layer with rates 0.1.

4 Experiments and Results

A series of experiments were carried out to establish the validity and stability of the proposed CNN architecture.

4.1 Set up

The dataset was split into disjoint training, validation and testing sets. First, the dataset was split into training and testing sets where 80% of the data was used for training and the remaining 20% for testing. The training set was then split into training and validation sets with ratios of 90% and 10% of the remaining training set respectively. The CNN model was trained with a batch size of 64 for 25 epochs. These parameters were set empirically.

4.2 Results & Discussion

On the training set, an accuracy of 99.8%, with only 2 symbols incorrectly classified was recorded. On the test set, results were slightly lower, with accuracy of 95.3%. In other words 23 symbols were incorrectly identified. Table 1 provides more details about performance across the different symbols and using three different metrics: Precision, Recall, and F1-Score.

Table 1: Performance across different symbols in the dataset

Symbol	Precision	Recall	F1-score	Symbol	Precision	Recall	F1-score
Control Valve	0.00	0.00	0.00	Control	1.00	1.00	1.00
Flange + Triangle	0.00	0.00	0.00	DB&BBV	1.00	1.00	1.00
Line Blindspacer	0.00	0.00	0.00	DB&BBV + Valve Check	1.00	0.90	0.95
Valve Gate Through Conduit	0.00	0.00	0.00	DB&BPV	1.00	0.95	0.98
Rupture Disc	0.33	1.00	0.50	Deluge	1.00	1.00	1.00
Valve Angle	0.33	1.00	0.50	ESDV Valve Ball	1.00	0.92	0.96
Valve Slab Gate	0.50	1.00	0.67	ESDV Valve Slab Gate	1.00	0.50	0.67
Valve Globe	0.57	1.00	0.73	Exit to Atmosphere	1.00	1.00	1.00
ESDV Valve Butterfly	0.67	1.00	0.80	Flange Single T-Shape	1.00	0.93	0.96
Control Valve Globe	0.80	0.57	0.67	Injector Point	1.00	0.62	0.77
Valve	0.80	1.00	0.89	Reducer	1.00	1.00	1.00
Flange Joint	0.85	1.00	0.92	Sensor	1.00	0.98	0.99
Arrowhead + Triangle	0.90	1.00	0.95	Spectacle Blind	1.00	1.00	1.00
Triangle	0.94	0.84	0.89	Valve Ball	1.00	0.97	0.99
Arrowhead	1.00	0.96	0.98	Valve Butterfly	1.00	1.00	1.00
Box	1.00	1.00	1.00	Valve Check	1.00	0.93	0.96
Continuity Label	1.00	1.00	1.00	Valve Plug	1.00	0.89	0.94

A closer look at the results, shows as expected that some of the minority class instances went completely undetected. For example, for the control valve

symbols which has only five instances in the whole dataset, the corresponding F1-score is zero. Such score can also be seen in Table 1 for the symbols the *'Flange + Triangle'* (17 instances in the whole dataset), the *'Line Blindspacer'* (4 instances only), *'Valve Gate Through Conduit'* with only 4 instances in the whole dataset. Conversely, well represented symbols in the dataset were correctly classified with relatively high precision and recall. For example, the *'Reducer'* F1-score is 1. Notice that 285 instances of reducers are present in the dataset. A similar performance can be observed for other majority class instances such as *'Sensor'* (392 instances), *'Valve Ball'* (173 instances in the dataset), and others.

These results are consistent with the literature and showed clearly that the learning algorithm tend to be biased toward majority class-instances. Despite this, it can be said that CNN performed extremely well on the testing set with an overall accuracy reaching 95.3%, and an average precision, recall, and F1-score of 0.785 ,0.822, and 0.784 respectively across all symbols in the dataset.

5 Conclusions

In this paper, we presented a new multiclass imbalanced dataset for the research community. The dataset represents a collection of symbols extracted from P&IDs. Despite the importance of processing and analysing engineering drawings, no such dataset exists in the public domain. We anticipate that donating this dataset to the research community will help researchers in the domain of machine learning and in particular imbalanced-class classification, and also research in the machine vision domain who are interested in processing and analysing engineering drawings. Future work will focus on handling this multiclass imbalanced problem, where advanced methods such as GANs and other data augmentation techniques might be utilised to improve the learning performance.

References

1. S. V. Ablameyko and S. Uchida. Recognition of engineering drawing entities: Review of approaches. *International Journal of Image and Graphics*, 07(04):709–733, 2007.
2. S. Ahmed, M. Liwicki, M. Weber, and A. Dengel. Automatic room detection and room labeling from architectural floor plans. In *2012 10th IAPR International Workshop on Document Analysis Systems*, pages 339–343, March 2012.
3. A. Ali-Gombe, E. Elyan, and C. Jayne. Multiple fake classes gan for data augmentation in face image dataset. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, July 2019.
4. Adamu Ali-Gombe and Eyad Elyan. Mfc-gan: Class-imbalanced dataset classification using multiple fake class generative adversarial network. *Neurocomputing*, 361:212 – 221, 2019.
5. E. Arroyo, A. Fay, M. Chioua, and M. Hoernicke. Integrating plant and process information as a basis for automated plant diagnosis tasks. In *Proceedings of the 2014 IEEE Emerging Technology and Factory Automation (ETFA)*, pages 1–8, Sep. 2014.

6. D. Blostein. General Diagram-Recognition Methodologies. In *Proceedings of the 1st International Conference on Graphics Recognition (GREC'95)*, pages 200–212, 1995.
7. C. F. Moreno-García, E. Elyan, C. Jayne. New trends on digitisation of complex engineering drawings. *Neural Computing and Applications*, Jun 2018.
8. A. K. Chhabra. Graphics Recognition Algorithms and Systems. In *Proceedings of the 2nd International Conference on Graphics Recognition (GREC'97)*, pages 244–252, 1997.
9. E. Elyan, C. F. Moreno-Garcia, and C. Jayne. Symbols classification in engineering drawings. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, July 2018.
10. Eyad Elyan and Mohamed Medhat Gaber. A fine-grained random forests using class decomposition: an application to medical diagnosis. *Neural Computing and Applications*, 27(8):2279–2288, 2016.
11. Eyad Elyan and Mohamed Medhat Gaber. A genetic algorithm approach to optimising random forests applied to class engineered data. *Information Sciences*, 384:220 – 234, 2017.
12. Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. A guide to deep learning in healthcare. *Nature Medicine*, 25(1):24–29, 2019.
13. Kim Nee Goh, Siti Rohkmah Mohd. Shukri, and Rofans Beleam Hilisebua Manao. Automatic assessment for engineering drawing. In Halimah Badioze Zaman, Peter Robinson, Patrick Olivier, Timothy K. Shih, and Sergio Velastin, editors, *Advances in Visual Informatics*, pages 497–507, Cham, 2013. Springer International Publishing.
14. Rafael C. Gonzalez and Richard E. Woods. *Digital image processing*. Prentice Hall, Upper Saddle River, N.J., 2008.
15. Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.
16. Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, and Tsuhan Chen. Recent advances in convolutional neural networks. *Pattern Recognition*, 77:354 – 377, 2018.
17. Sean D. Holcomb, William K. Porter, Shaun V. Ault, Guifen Mao, and Jin Wang. Overview on deepmind and its alphago zero ai. In *Proceedings of the 2018 International Conference on Big Data and Education*, ICBDE '18, pages 67–71, New York, NY, USA, 2018. ACM.
18. C Howie, J Kunz, T Binford, T Chen, and K.H Law. Computer interpretation of process and instrumentation drawings. *Advances in Engineering Software*, 29(7):563 – 570, 1998.
19. T. Kanungo, R. M. Haralick, and D. Dori. Understanding Engineering Drawings: A Survey. In *Proceedings of the 1st International Conference on Graphics Recognition (GREC'95)*, pages 119–130, 1995.
20. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, May 2017.
21. C. R. Kulkarni and A. B. Barbadekar. Text Detection and Recognition: A Review. *International Research Journal of Engineering and Technology (IRJET)*, 4(6):179–185, 2017.
22. Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, Nov 1998.

23. Y. Lu. Machine printed character segmentation - An overview. *Pattern Recognition*, 28(1):67–80, 1995.
24. C. F. Moreno-Garcia and E. Elyan. Digitisation of assets from the oil gas industry: Challenges and opportunities. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 7, pages 2–5, Sep. 2019.
25. C. F Moreno-García, E Elyan, and C Jayne. Heuristics-Based Detection to Improve Text / Graphics Segmentation in Complex Engineering Drawings. In *Engineering Applications of Neural Networks*, volume CCIS 744, pages 87–98, 2017.
26. S. Mori, C. Y. Suen, and K. Yamamoto. Historical Review of OCR Research and Development. *Proceedings of the IEEE*, 80(7):1029–1058, 1992.
27. U. Park and A. K. Jain. Face matching and retrieval using soft biometrics. *IEEE Transactions on Information Forensics and Security*, 5(3):406–415, Sep. 2010.
28. N. Shone, T. N. Ngoc, V. D. Phai, and Q. Shi. A deep learning approach to network intrusion detection. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2(1):41–50, Feb 2018.
29. Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.
30. C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, June 2015.
31. Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, June 2014.
32. Karl Tombre, Salvatore Tabbone, Bart Lamiroy, and Philippe Dosch. Text/Graphics Separation Revisited. In *Document Analysis Systems*, volume 2423, pages 200–211, 2002.
33. P. Vaxiviere and K. Tombre. Celesstin: Cad conversion of mechanical drawings. *Computer*, 25(7):46–54, July 1992.
34. Pattaramon Vuttipittayamongkol and Eyad Elyan. Neighbourhood-based undersampling approach for handling imbalanced and overlapped data. *Information Sciences*, 509:47 – 70, 2020.
35. Pattaramon Vuttipittayamongkol, Eyad Elyan, Andrei Petrovski, and Chrisina Jayne. Overlap-based undersampling for improving imbalanced data classification. In Hujun Yin, David Camacho, Paulo Novais, and Antonio J. Tallón-Ballesteros, editors, *Intelligent Data Engineering and Automated Learning – IDEAL 2018*, pages 689–697, Cham, 2018. Springer International Publishing.
36. Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California, June 2016. Association for Computational Linguistics.
37. D. Zhang and G. Lu. Review of shape representation and description techniques. *Pattern Recognition*, 37(1):1–19, 2004.