Using Natural Language Processing to uncover main topics in defect recognition literature

Bernabei M.*, Colabianchi S.*, Costantino F.*, Patriarca R.*

** Department of Mechanical and Aerospace Engineering, University of Rome "La Sapienza", Via Eudossiana, 18, 00184 Rome, Italy (bernabei.1708149@studenti.uniroma1.it, silvia.colabianchi@uniroma1.it, francesco.costantino@uniroma1.it, riccardo.patriarca@uniroma1.it)*

Abstract: The issue of defect detection is particularly important namely in plant engineering, where it is crucial to ensure high-quality production by minimizing the number of defective parts. In the last years, the interest in the subject has grown a lot and the methods and approaches proposed for defect recognition are multiple. Therefore, when dealing with defect recognition researchers are faced with an increasing number of articles that slows them down in identifying the set of articles of their interest. This work aims to provide a baseline classification of articles based on emerging issues such as the investigated material, the production typology in which the material is included, and the type of analysis to be effected. For these reasons, the paper proposes an automatic solution based on text mining techniques. Specifically, the study applies Natural Language Processing (NLP) to articles' titles, abstracts, and keywords using two different approaches: K-Means clustering algorithm and Latent Dirichlet Allocation (LDA). K-Means is used to cluster the collection of documents into related groups based on the contents of the particular documents. LDA instead is used to classify the papers using the concept of topic modeling. Articles have been collected from Scopus database. The scope of the research is limited to journal and conference articles, published in English excluding articles classified as reviews, as well as book chapters, books, notes, erratum.

Keywords: K-Means, LDA, information retrieval (IR); clustering; manufacturing

1. Introduction

Over the last decade, the topic of defect recognition has gained much attention and a considerable amount of studies about different approaches to detect defects in the production context has been accumulated. In 2010 a Scopus search for "defect recognition" in production contexts yielded 194 papers, while in 2021 the same search retrieved more than twice that amount (542 papers). This is also reflected in today's industrial business environment. The increased quality demand and at the same time the need to reduce both cost and delivery time forces organizations to define and employ adequate business processes to meet these demands. An adequate verification and validation process encompassing defect recognition and prevention is an indispensable contribution to ensure high-quality production (Quatrini *et al.*, 2020). However, nowadays it has become more challenging to apply systematic monitoring and detection of defects occurrence due to the increasing complexity of both products and production systems (Van Moll *et al.*, 2002). Traditional inspection methods have become costly, slow, and ineffective. For this reason, companies have increased their investments in sophisticated strategies and tools to identify defected products (Psarommatis *et al.*, 2020). Also in the literature today we find many innovative methods and approaches for defect recognition. Multiple studies have focused on a specific type of materials, such as fabric (Zhou *et al.*, 2021), steel (Luo *et al.*, 2020); a group of materials such as web materials (Bulnes *et al.*, 2016); a specific part of the production process such as welding (Zhang *et al.*, 2021); a particular product e.g. semiconductors (Diachenko *et al.*, 2016) or solar cells (Ying *et al.*, 2019). Other studies have reviewed defect detection approaches such as deep learning (Dizaji and Harris, 2019), inspection, thermographic image processing (Wang and Zhao, 2018), ultrasonic techniques and artificial neural network (Yuan *et al.*, 2020), visual-based detection and classification (Czimmermann *et al.*, 2020) which describe a dedicated group of techniques applicable to different materials. As can be seen, the theme of defects recognition is vast and the choice of the most suitable method depends on several aspects, such as the investigated material, the production typology in which the material is included, the type of analysis to be effected, etc. (Luo *et al.*, 2020). This survey is not intended to be specific to one of the aforementioned aspects but is intended to provide an overview of them to allow researchers to get a global idea on the topic of defect recognition and facilitate an exchange of knowledge. A complete analysis of the different types of defects, materials, and technologies used in the various production contexts would be difficult to perform by manual means (e.g. a Preferred Reporting Items for Systematic Reviews and Meta-Analyses based approach). For this purpose, we developed and discussed a text mining tool to classify defects recognition literature. Our tool automates the process by extracting relevant scientific data in published literature, highlighting the main research topics presented, and rapidly identifying documents that may be relevant for a specific production process. For the development of this tool, we drew on techniques of NLP and text mining. NPL techniques imply the automatic processing of written or spoken information of a natural language by means of an electronic calculator. Generally speaking, NLP belongs to a field of studies where computer science, artificial intelligence and linguistics are combined together, thus producing a programming method aimed at processing large quantities of data with various objectives, ranging from information extraction through different text processing methods, to emotion detection, i.e. Sentimental Analysis. NLP enables retrieval, analysis, and information condensation mechanisms, while also recognizing models and patterns, labeling large databases, and accelerating operations for the display of hidden structures by using algorithms and analytical methods. In this context, written text data represent a crucial input. Text mining techniques have shown an increase in their use, and it is also interesting to note a growing interest in the industrial sector. NLP has been applied, for example, to industrial risk management. In the study proposed by Kwayu *et al.* (2021), NLP has been used for the analysis of incident and accident narratives, in the work of Li *et al.* (2021) instead text mining was performed to extract additional information from the narratives in reports related to natural gas distribution pipelines. NLP has been also used to improve complex production, such as the production of microchips, analyzing documents relative to the production process for integrating lessons learned (Abu Rasheed *et al.*, 2021). Differently, Wang and Hsu (2020) have proposed an approach for exploring technological trends in smart manufacturing analysing text contained in patent data through NLP and topic modelling algorithms. Similarly, Barnewold and Lottermoser (2020) have used text-mining techniques to identify relevant digital technologies pursued in the mining industry and relations between significant digital trends. This work falls into this final group of NLP techniques. The one part of NLP aimed at extracting concepts, areas of investigation, and research topics from a large number of documents collected within the existing literature by means of text analysis logic (Chiarello et al., 2019; Weißer et al., 2020). To our knowledge, no study reveals the main themes in defect recognition literature. Moreover as suggested also by Weißer (2020) there is a lack of efficient methods to ensure objectivity in the selection and filtering of the articles during the phase of literature analysis. Identifying main topics will also help researchers to better clarify the evaluation criteria for the exclusion or retention of a paper. Finally, as mentioned before, a last innovative aspect is the comparison of two different approaches: LDA and K-Means. The first one, K-Means, partitions the initial dataset in disjointed K-clusters (Weißer *et al.*, 2020), while the second one, LDA, uses the initial dataset to create a mixture of topics, to which each document is subsequently assigned according to probabilistic logic models (Chiarello *et al.*, 2019). In fact, a single document can be assigned, with differing percentages, to more than one topic, thus providing more realistic and elucidative results. The sections of this paper are organized as follows. Section 2 describes the techniques and the algorithms employed. Section 3 discuss the results of the survey and presents a comparison between the two approaches used. Finally Section 4 presents the main encountered issues as well as the future development lines for the continuous improvement of the proposed method.

2. Proposed approach

The pre-processing phase is based on NLP techniques and it is structured in 4 steps: tokenization, stop word and sparse terms removal, stemming and lemmatization, identification of n-grams, and Term Frequency and Inverse Document Frequency (TFIDF) vectorization. In the last step, K-Means and LDA are applied on pre-processed data and used to automatically break down articles corpora into distinct topical groups. The method overview is shown in Figure 1.

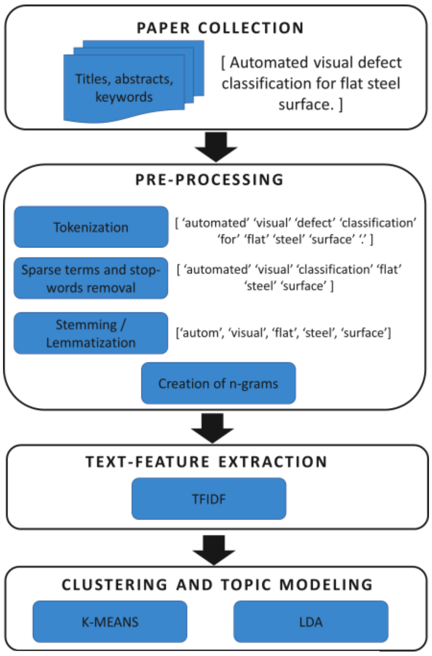

Figure 1 – Proposed method steps

2.1. Paper collection

The initial step is collecting a set of research articles gathered from large databases of peer-reviewed literature. The approach is applied to articles' titles, keywords, and abstracts. Before going to the following step, any duplicate article must be removed.

2.2. Pre-Processing

In this second step, a set of data pre-processing activities is conducted. Data pre-processing is a method of data mining that involves transforming raw data into a reasonable format for topic modelling and cluster analysis. The pre-processing stage is develop using NLTK library in Python. For the analysis of text, data collected the following NLP steps must be pursued.

Tokenization

Since documents are unstructured information, they must be divided into linguistic units. The process of splitting a phrase, sentence, paragraph, or entire text document into smaller units is called tokenization and these pieces are called tokens. There are different tokenization strategies, in this article white spaces between words are considered as a separator. In this step also punctuation is removed.

Stop-words and sparse terms removal

Stop-words are commonly used words (e.g. articles, prepositions, etc.) that occur frequently in article corpora but do not carry any meaning on their own. Depending on the domain and language there will be a different set of stop-word. Moreover, for better performance, it is recommended to remove search terms used in the query that may bias the clustering results. Sparse terms are instead terms that occur in less than 1% of the documents.

Stemming / Lemmatization

Stemming is a process of reducing the words to their root form. Lemmatization instead is a more complex version of stemming which uses part of speech (POS) tagging for each word, (e.g. verb, noun, etc). First, each token is tagged with a POS, second lemmatization applies a different stemming rule to each token depending on the tagged POS. This allows diminishing the lexical sparsity of the corpus since two words that have the same root result in the same output.

Ngrams

N-grams are n words frequently occurring together in the document. N-grams estimate the probability of the next item in a word sequence calculating the occurrence of the next word with respect to the previous one. In this analysis are consider b-grams - two-word sequence of words – and tri-grams – a three-word sequence of words.

2.3. Text Features Extraction

TFIDF vectorization

The last step in the NLP pipeline is TFIDF vectorization. The TFIDF measure reflects how important a word is to a document in a collection of documents and it consists of two parts. The first part (TF) is term frequency which counts how many times a word appeared in each document considered. The second part (IDF) is inverse document frequency which is responsible for reducing the weights of words that occur frequently and increase the weights of words that occur rarely (Hari Prasath *et al.,* 2016).

2.4. Processing

The approach proposed uses both K-Means and LDA algorithms.

K-Means

K-means is an unsupervised clustering algorithm that partitions the N documents in K, deliberately selected, disjoint clusters. Once the number of clusters is set, the algorithm randomly selects K samples for the centroids. Then each observation is assigned to a cluster to minimize the within-cluster sum of squares. Next, the mean of the clustered observations is calculated and used as the new cluster centroid. Then, observations are reassigned to clusters, and centroids recalculated in an iterative process until the algorithm reaches convergence.

The optimum number of clusters is chosen using the silhouette score. The silhouette is a measure of how close each point in one cluster is to points in the neighbouring clusters. The silhouette coefficient varies between -1 and 1. A values near 0 indicate overlapping clusters. Whereas, a value close to -1 means that the value is assigned to the wrong cluster (Weißer *et al.*, 2020).

LDA

LDA is one of the most used topics modelling algorithms and assigns a document to a mixture of topics (Chiarello *et al.*, 2019). Specifically, LDA is a generative probabilistic model for counting the occurrences of words and collections of discrete data from text corpora. LDA assumes that each topic is a mixture over an underlying set of words, and each document is a mixture of over a set of topic probabilities. It is characterized by two main parameters defined as follows: (a) Alpha parameter, which is a concentration parameter and represents document-topic density. (b) Beta parameter which represents topic-word density. The optimum number of topics is chosen using perplexity and coherence value using perplexity score, a statistical measure of how well a probability model predicts a sample and topic coherence score which measure the degree of semantic similarity between high scoring words in the topic value (Lau, Baldwin and Newman, 2013). The goal is to choose the number of topics that minimize the perplexity and maximize the coherence value compared to other numbers of topics.

3. Results and discussion

3.1 Results from paper collection

The initial set of papers was collected by searching on Scopus database "defect recognition" in Titles, abstracts and keywords, limiting the query to this subject areas: engineering, computer science, mathematics, chemical. As a result of the above-mentioned selection, 542 documents on defect detection were retrieved.

3.2 Results from K-Means

The first analysis was performed by employing the K-Means algorithm. For the purpose of this analysis, the number of K-clusters was chosen according to the results obtained from the Silhouette analysis. The results obtained with the Silhouette analysis indicated that the most suitable number of clusters (K) for the analysis under consideration was 4. In fact, by setting K=4 it was possible to ensure a fair distance between the clusters. Moreover, all the clusters had comparable dimensions and none of them showed a score lower than the Silhouette mean (0.650). The clusters obtained by setting K=4 are presented in Table 1. Firstly, Cluster 0 shows mainly topics related to defect detection with regard to the involved techniques, the most frequent being those concerning the inspection, measurement and use of deep learning networks. Moreover, the techniques contained in Cluster 0, due to the ample size of the cluster, can be employed in various application fields, such as in defect detection through convolutional neural networks (Jun *et al.*, 2021), in fibers by means of unsupervised methods (Siegmund *et al.*, 2020), in the photolithography process through neural networks (Chen, Su and Chen, 2009), in metallurgic parts through ultrasonic inspection techniques (Turó *et al.*, 2013), etc.

Table 1 – K-Means clusters outline

#Cluster	Cluster's key terms	#Elements in the cluster
0	Inspection, Signal, Production, Network, Measurement, Pattern	299
1	Wood, Rust, Color, Log, Coat, Steel-Bridge	49
2	Weld, Radiographs, Geometrical, CNN, Stage, Class	35
3	Surface, Classifier, Segment, Rate, Extract, Accuracy	159

Secondly, Cluster 1 presents the topic of defect detection mainly in terms of surface and image analysis in correlation with specific materials, i.e. wood and steel. In particular, the presented techniques concern wood image analysis for surface (Yang *et al.*, 2020) and internal (Zhu and Beex, 1994) defect detection, as well as surface analysis of nickel foam (Cao, Li and Qiao, 2020), etc. Furthermore, Cluster 2 contains documents related to defect detection in geometrical and radiographic analyses, with a focus on the techniques for welding regions. For instance, the welding's quality can be analysed by means of region division techniques (Zhang *et al.*, 2021), ultrasound registration techniques for positioning and dimension (Serrano *et al.,* 1999), and in X-ray welding processes, through feature extraction for defect detection with grey enhancement algorithms (Kuang *et al.,* 2016). Finally, Cluster 3 comprises defect detection techniques concerning classification, segmentation, and accuracy. The topics included in this cluster are, for example, feature collection for accurate defect classification and discrimination, and two-phase defect classification, consisting of an image segmentation phase and a subsequent feature extraction phase (Wan *et al.*, 2019), which can have several applications, namely in the textile industry (Xia *et al.*, 2017).

3.3 Results from LDA

The second analysis was performed by implementing the LDA algorithm. The number of topics (6) was chosen according to the maximum coherence score obtained. The topics' top terms obtained are shown in Table 2. These terms should be analyzed together with the probability distribution over the K topics for each paper. Each term paper is indeed attributed to a particular topic with probability given by this distribution. Topic 1 is mainly focused on researches related to detecting defects on different surfaces. The most influential papers within the topic focus on automated visual inspection of textile surfaces (Xia *et al.*, 2017; Zhou *et al.*, 2021) and the detection of defects linked to the casting manufacturing process (e.g. Brunke *et al.*, 2013; Yu *et al.*, 2020). Still in this topic, but registering a lower percentage of contribution to the topic, are those documents that explore defects recognition in electronic packages especially flip-chip devices (e.g. Liao *et al.*, 2015; Fan *et al.*, 2016). Extracts from topic 2 show that a high percentage of contribution is assigned to papers presenting computer vision inspection approaches. A significant group of studies focuses on the use of X-Ray topography for investigations of structural defects (Kuang, Xie and Zhao, 2016) and near-infrared (NIR) images for the investigations of crystalline perfection and metallic impurity (Cui *et al.*, 2020). Topic 3 can be mainly attributed to models trained for the recognition and evaluation of defects associated related to the construction industry (e.g. Di *et al.*, 2018; Zhao *et al.*, 2021). However, within Topic 3 a subset of articles related to defect detection in steel products is identified. Articles related to steel surface defect classification (Luo *et al.*, 2020), steel plates (Xu *et al.,*2010), or steel strips (Zhang *et al.*, 2020). In Topic 4, the most influential papers are those dealing with issues of defect recognition through ray inspection techniques applied to solder joint (e.g. Gao, Zhang and Mi, 2019) or rust defects investigated by image color analysis (Lee *et al.,* 2006). A subgroup of less influential papers using Convolutional Neural Network (CNN) is also identified in this topic. For example, Dizaji and Harris (2019) use CNN trained on 3D images applied to internal defects of workpiece products, Huang et al. (2019) use it to solve the railway infrastructure defect detection problem, or Shen & Yu (2019) apply CNN to recognize wafer map defects. Topic 5 focuses mostly on defects related to wood products and woodworking in general.

Table 2: LDA topics outline.

#Topic	Topic identification terms
1	Inspect, product, textil, materi, autom, requir, measur, manufactur, object, crack, cast, develop, ray, equip, comput.
2	Weld, surfac, detect, qualiti, region, develop, real_tim, optic, edg, order, evalu, discuss, laser, machine_vis, nois.
3	Fabric, surfac, recogn, pattern, type, set, select, effect, extract, signal, failur, shape, network, studi, classifi.
4	Segment, color, rust, inspect, problem, solder_joint, digit, surfac, perform, morpholog, accuraci, ray_inspect, develop, object, machine_vis.
5	Wood, pattern, log, classifi, textur, hardwood_log, signal, characterist, robust, accuraci, grade, extract, insul, filter, statist.
6	Inspect, knowledg, design, industri, structur, system, scan, generat, measur, modul, board, stage, imag, develop, present.

The most influential papers discuss the pre-process and analysis of wood images, for example, to extract the wood defect contour (Y. Yang et al., 2020), to industrialize the wood defects recognition (Mu et al., 2015), or modelling techniques applied to wood-grain texture analysis of CT images of hardwood logs (Zhu & Beex, 1994). Finally, in Topic 6 we find again a cluster related to CNN for defects' recognition that in this scenario brings a major contribution to the topic. CNN is applied for example to Metal Additive Manufacturing to inspect qualitative aspect such as good quality, crack, and porosity (Cui et al., 2020), or used to detect defects in PV module cells (Ying *et al.*, 2019), gray cloth (Kuo et al., 2008) or semiconductors (Patel, Bonam and Oberai, 2019). Finally, it is interesting to note that within the latter topic there is a large subset of articles dealing with defect detection and classification in the semiconductor industry (e.g. Diachenko et al., 2016; Patel, Bonam and Oberai, 2019).

3.4. Algorithm comparison

Overall, the comparison of the results obtained with the K-Means and LDA algorithms indicates that the LDA algorithm reveals the presence of subsets within the clusters identified through the K-Means technique. For instance, the K-Means analysis placed documents (Shu *et al.,* 2021) and (Jun *et al.*, 2021) within Cluster 0, since they both concern defect detection through surface analysis. Contrastingly, according to the LDA classification, the first topics assigned to the two above-mentioned documents were different. In fact, the first topic assigned to document (Shu *et al.,*, 2021), which relates to surface inspection in LED chips, was Topic 4, and the first topic assigned to document (Jun *et al.*, 2021), which concerns textile surface inspection was Topic 3. For the aforementioned reasons, the LDA method seems to be able to provide more pragmatical results for the investigation objective, i.e. to face from a methodological point of view an industrial issue related to defect recognition. In fact, the perspective is to understand in general which are the methodological trends, and eventually to be able to notice contaminations or common trends within apparently distant sectors. However, to make the best use of both methods' results it is always necessary to consider what the final analysis objective is, since the two algorithms provide a different type of information, precisely because of the logic on which they are based.

4. Conclusion

The vast amount of academic articles has led researchers to struggle in classifying and labelling studies of their interest. For this reason, this paper helps researchers to understand key trends within defect recognition in the industrial context. Specifically, the approach applied NLP to articles' metadata along with K-Means and LDA algorithms to extract useful insights and major research topics. Results of both K-Means and LDA have proven to be very satisfactory in identifying the different aspects, materials, and dimensions for which much attention is paid to the issue of defect recognition. More common techniques emerged within certain industries, such as textiles or metallurgy. Shared techniques within multiple sectors also emerged. This underscores the possibility of sharing the same techniques across industries, and for this reason, it is increasingly important to have a broader view on the topic. Altogether, the method presented allows an initial classification of a large set of papers, which typically must be faced when new and broad themes are investigated. In particular, the method highlights the presence of different techniques and subject areas. This organized overview of the different aspects of defect recognition can offer researchers the possibility to investigate only those papers related to the clusters of interest. Indeed, the analysis presented is not exhaustive and needs further elaboration. In particular, a more detailed analysis of the various articles assigned to clusters and topics is necessary to have a more detailed review on the subject of defects recognition, which will be future research. Moreover, this study is limited in the pre-processing phase, specifically in the identification of sparse terms and query-related terms to be removed. At this stage, our tool still requires "human-in-the-loop" to identify academic terms related to the query that could cause bias in the clusters' creation. Future research will need to consider a fully automated solution to identify terms to be removed. Finally, K-Means results, despite a good silhouette metric, resulted in having a more unbalanced cluster that collected several topics within it, thus losing some information that could have been interesting. Future research should test other methods such as, the hierarchical clustering method which might help in the identification of the optimum number of clusters and offers good interpretable and informative results despite requiring a higher computational effort. For this reason, it would also be interesting to merge these two algorithms and test them on the dataset.

References

Abu Rasheed, H. *et al.* (2021) 'A Text Extraction-Based Smart Knowledge Graph Composition for Integrating Lessons Learned During the Microchip Design', in *Advances in Intelligent Systems and Computing*. Springer, pp. 594–610. doi: 10.1007/978-3-030-55187-2_43.

Barnewold, L. and Lottermoser, B. G. (2020) 'Identification of digital technologies and digitalisation trends in the mining industry', *International Journal of Mining Science and Technology*. China University of Mining and Technology, 30(6), pp. 747–757. doi: 10.1016/j.ijmst.2020.07.003.

Brunke, O. *et al.* (2013) 'A new concept for high-speed Atline and Inlinect for up to 100% mass production process control', in *12th International Conference of the Slovenian Society for Non-Destructive Testing: Application of Contemporary Non-Destructive Testing in Engineering, ICNDT 2013 - Conference Proceedings*, pp. 265–270. Available at: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84930913009&partnerID=40&md5=eb4b77730d3ca1bf9af2b38636eb49ec.

Bulnes, F. G. *et al.* (2016) 'An efficient method for defect detection during the manufacturing of web materials', *Journal of Intelligent Manufacturing*. Springer New York LLC, 27(2), pp. 431–445. doi: 10.1007/s10845-014-0876-9.

Cao, B.-F., Li, J.-Q. and Qiao, N.-S. (2020) 'Nickel foam surface defect detection based on spatial-frequency multi-scale MB-LBP', *Soft Computing*, 24(8), pp. 5949–5957. doi:

10.1007/s00500-019-04513-2.

Chen, L.-F., Su, C.-T. and Chen, M.-H. (2009) 'A neural-network approach for defect recognition in TFT-LCD photolithography process', *IEEE Transactions on Electronics Packaging Manufacturing*, 32(1), pp. 1–8. doi: 10.1109/TEPM.2008.926117.

Chiarello, F. *et al.* (2019) 'A text mining based map of engineering design: Topics and their trajectories over time', in *Proceedings of the International Conference on Engineering Design, ICED*. Cambridge University Press, pp. 2765–2774. doi: 10.1017/dsi.2019.283.

Cui, W. *et al.* (2020) 'Metal additive manufacturing parts inspection using convolutional neural network', *Applied Sciences (Switzerland)*, 10(2). doi: 10.3390/app10020545.

Czimmermann, T. *et al.* (2020) 'Visual-based defect detection and classification approaches for industrial applications—A SURVEY', *Sensors (Switzerland)*. MDPI AG, p. 1459. doi: 10.3390/s20051459.

Di, W.-G. *et al.* (2018) 'Recognition of Bolt Quality Base on Elman Neural Network by Ant Colony Optimization Algorithm', in *Proceedings - International Conference on Machine Learning and Cybernetics*, pp. 210–214. doi: 10.1109/ICMLC.2018.8527051.

Diachenko, L. *et al.* (2016) 'New computer system for recognizing micro- and nano-sized objects in semiconductors and colloidal solutions', *Journal of Nano- and Electronic Physics*, 8(4). doi: 10.21272/jnep.8(4(2)).04060.

Dizaji, M. S. and Harris, D. K. (2019) '3D InspectionNet: A deep 3D convolutional neural networks based approach for 3D defect detection on concrete columns', in *Proceedings of SPIE - The International Society for Optical Engineering*. doi: 10.1117/12.2514387.

Fan, M. *et al.* (2016) 'Defect inspection of solder bumps using the scanning acoustic microscopy and fuzzy SVM algorithm', *Microelectronics Reliability*, 65, pp. 192–197. doi: 10.1016/j.microrel.2016.08.010.

Gao, S., Zhang, H. and Mi, H. (2019) 'Solder Joint Defect Detection Based on Image Segmentation and Deep Learning', in *ICSIDP 2019 - IEEE International Conference on Signal, Information and Data Processing 2019*. doi: 10.1109/ICSIDP47821.2019.9173443.

Hari Prasath, B., Karthikeyan, S. and Mary Valantina, G. (2016) 'Computerized highway defects recognition and classification system', *International Journal of Pharmacy and Technology*, 8(1), pp. 11038–11048. Available at: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84963567592&partnerID=40&md5=980b0e0eb012dbb62f932d954682ec21.

Huang, H. *et al.* (2019) 'Railway Infrastructure Defects Recognition using Fine-grained Deep Convolutional Neural Networks', in *2018 International Conference on Digital Image Computing: Techniques and Applications, DICTA 2018*. doi: 10.1109/DICTA.2018.8615868.

Jun, X. *et al.* (2021) 'Fabric defect detection based on a deep convolutional neural network using a two-stage strategy', *Textile Research Journal*, 91(1–2), pp. 130–142. doi: 10.1177/0040517520935984.

Kuang, P., Xie, Z. and Zhao, Q. (2016) 'Weld region extraction in radiographic image based on template match and gray enhancement', in *2015 12th International Computer Conference on Wavelet Active Media Technology and Information Processing, ICCWAMTIP 2015*, pp. 224–227. doi: 10.1109/ICCWAMTIP.2015.7493980.

Kuo, C.-F. J. *et al.* (2008) 'Intelligence control of on-line dynamic gray cloth inspecting machine system module design. II. Defects inspecting module design', *Fibers and Polymers*, 9(6), pp. 768–775. doi: 10.1007/s12221-008-0120-3.

Kwayu, K. M. *et al.* (2021) 'Discovering latent themes in traffic fatal crash narratives using text mining analytics and network topology', *Accident Analysis and Prevention*. Elsevier Ltd, 150. doi: 10.1016/j.aap.2020.105899.

Lau, J. H., Baldwin, T. and Newman, D. (2013) 'On collocations and topic models', *ACM Transactions on Speech and Language Processing*, 10(3). doi: 10.1145/2483969.2483972.

Lee, S., Chang, L.-M. and Skibniewski, M. (2006) 'Automated recognition of surface defects using digital color image processing', *Automation in Construction*, 15(4), pp. 540–549. doi: 10.1016/j.autcon.2005.08.001.

Li, M.-X. and Wan, Y.-J. (2016) 'Research on the solder joint image segmentation based on the improved spatial fuzzy C means algorithm', in *2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery, ICNC-FSKD 2016*, pp. 1940–1944. doi: 10.1109/FSKD.2016.7603476.

Li, X. *et al.* (2021) 'Severity of emergency natural gas distribution pipeline incidents: Application of an integrated spatio-temporal approach fused with text mining', *Journal of Loss Prevention in the Process Industries*. Elsevier Ltd, 69. doi: 10.1016/j.jlp.2020.104383.

Liao, G. *et al.* (2015) 'Using RBF networks for detection and prediction of flip chip with missing bumps', *Microelectronics Reliability*, 55(12), pp. 2817–2825. doi: 10.1016/j.microrel.2015.09.030.

Luo, Q. *et al.* (2020) 'Automated Visual Defect Classification for Flat Steel Surface: A Survey', *IEEE Transactions on Instrumentation and Measurement*, 69(12), pp. 9329–9349. doi: 10.1109/TIM.2020.3030167.

Van Moll, J. H. *et al.* (2002) 'The importance of life cycle modeling to defect detection and prevention', in *Proceedings - 10th International Workshop on Software Technology and Engineering Practice, STEP 2002*. Institute of Electrical and Electronics Engineers Inc., pp. 144–155. doi: 10.1109/STEP.2002.1267624.

Mu, H. *et al.* (2015) 'Wood defects recognition based on fuzzy BP neural network', *International Journal of Smart Home*, 9(5), pp. 143–152. doi: 10.14257/ijsh.2015.9.5.14.

Patel, D., Bonam, R. and Oberai, A. A. (2019)

'Engineering neural networks for improved defect detection and classification', in *Proceedings of SPIE - The International Society for Optical Engineering*. doi: 10.1117/12.2515065.

Pluska, M. *et al.* (2012) 'Defect detection in semiconductor layers with built-in electric field with the use of cathodoluminescence', *Physica B: Condensed Matter*, 407(15), pp. 2854–2857. doi: 10.1016/j.physb.2011.08.095.

Psarommatis, F. *et al.* (2020) 'Zero defect manufacturing: state-of-the-art review, shortcomings and future directions in research', *International Journal of Production Research*. Taylor and Francis Ltd., pp. 1–17. doi: 10.1080/00207543.2019.1605228.

Quatrini, E. *et al.* (2020) 'Condition-based maintenance-An extensive literature review', *Machines*, 8(2). doi: 10.3390/MACHINES8020031.

Serrano, I., Lazaro, A. and Oria, J. P. (1999) 'Ultrasonic inspection of foundry pieces applying wavelet transform analysis', in *IEEE International Symposium on Intelligent Control - Proceedings*, pp. 375–380. doi: 10.1109/isic.1999.796684.

Shu, Y., Li, B. and Lin, H. (2021) 'Quality safety monitoring of LED chips using deep learning-based vision inspection methods', *Measurement: Journal of the International Measurement Confederation*, 168. doi: 10.1016/j.measurement.2020.108123.

Siegmund, D. *et al.* (2020) 'Detection of Fiber Defects Using Keypoints and Deep Learning', *International Journal of Pattern Recognition and Artificial Intelligence*. doi: 10.1142/S0218001421500166.

Sun, X. *et al.* (2016) 'The study of nondestructive testing of rock bolts based on PNN and wavelet packet', in *Proceedings of 2015 7th International Conference on Modelling, Identification and Control, ICMIC 2015*. doi: 10.1109/ICMIC.2015.7409489.

Turó, A. *et al.* (2013) 'Ultrasonic inspection system for powder metallurgy parts', *Measurement: Journal of the International Measurement Confederation*, 46(3), pp. 1101–1108. doi: 10.1016/j.measurement.2012.10.016.

Wan, D. *et al.* (2019) 'A Partial Discharge Pattern Recognition Method Based on Optimal Setting of characteristic Parameters and Classifiers', in *2019 3rd IEEE Conference on Energy Internet and Energy System Integration: Ubiquitous Energy Network Connecting Everything, EI2 2019*, pp. 2874–2877. doi: 10.1109/EI247390.2019.9061917.

Wang, J. and Hsu, C. C. (2020) 'A topic-based patent analytics approach for exploring technological trends in smart manufacturing', *Journal of Manufacturing Technology Management*. Emerald Group Holdings Ltd., 32(1), pp. 110–135. doi: 10.1108/JMTM-03-2020-0106.

Wang, J. and Zhao, C. (2018) 'Visual defect recognition and location for pulsed thermography images based on defect-background contrast analysis', in *Proceedings - 2018 33rd Youth Academic Annual Conference of Chinese Association of Automation, YAC 2018*, pp. 1106–1111. doi: 10.1109/YAC.2018.8406536.

Weißer, T. *et al.* (2020) 'A clustering approach for topic filtering within systematic literature reviews', *MethodsX*, 7. doi: 10.1016/j.mex.2020.100831.

Xia, D. *et al.* (2017) 'Warp-knitted fabric defect segmentation based on non-subsampled Contourlet transform', *Journal of the Textile Institute*, 108(2), pp. 239–245. doi: 10.1080/00405000.2016.1161700.

Xu, K., Wu, X. and Ai, Y. (2010) 'Surface defect recognition of hot rolled steel plates based on amplitude spectrum partitioning', in *Proceedings of the 10th International Conference on Steel Rolling*. Available at: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84867636028&partnerID=40&md5=688a63e5957866deedb7eb51ade4c669.

Yang, Y. *et al.* (2020) 'Wood defect detection based on depth extreme learning machine', *Applied Sciences (Switzerland)*, 10(21), pp. 1–14. doi: 10.3390/app10217488.

Ying, Z. *et al.* (2019) 'Automatic Detection of Photovoltaic Module Cells using Multi-Channel Convolutional Neural Network', in *Proceedings 2018 Chinese Automation Congress, CAC 2018*, pp. 3571–3576. doi: 10.1109/CAC.2018.8623258.

Yu, H. *et al.* (2020) 'Adaptive depth and receptive field selection network for defect semantic segmentation on castings X-rays', *NDT and E International*, 116. doi: 10.1016/j.ndteint.2020.102345.

Yuan, M. *et al.* (2020) 'Automatic recognition and positioning of wheel defects in ultrasonic B-scan image using artificial neural network and image processing', *Journal of Testing and Evaluation*, 48(1). doi: 10.1520/JTE20180545.

Zhang, S. *et al.* (2021) 'Far-sided defect recognition of FRP sandwich structures based on local defect resonance', *Journal of Sandwich Structures and Materials*, 23(2), pp. 568–579. doi: 10.1177/1099636219840250.

Zhang, Z. F. *et al.* (2020) 'Steel strip surface inspection through the combination of feature selection and multiclass classifiers', *Engineering Computations (Swansea, Wales)*. doi: 10.1108/EC-11-2019-0502.

Zhao, J. *et al.* (2021) 'Defect Recognition in Concrete Ultrasonic Detection Based on Wavelet Packet Transform and Stochastic Configuration Networks', *IEEE Access*, 9, pp. 9284–9295. doi: 10.1109/ACCESS.2021.3049448.

Zhou, T. *et al.* (2021) 'EDDs: A series of Efficient Defect Detectors for fabric quality inspection', *Measurement: Journal of the International Measurement Confederation*, 172. doi: 10.1016/j.measurement.2020.108885.

Zhu, D. and Beex, A. A. L. (1994) 'Robust spatial autoregressive modeling for hardwood log inspection', *Journal of Visual Communication and Image Representation*, 5(1), pp. 41–51. doi: 10.1006/jvci.1994.1004.