BI (BUSINESS INTELLIGENCE)-ТИЗИМЛАРИДА МАЪЛУМОТЛАРГА ДАСТЛАБКИ ИШЛОВ БЕРИШ ТАСНИФЛАРИ

Бабомурадов Озод Жўраевич
Жиззах шаҳридаги Қозон (Волгабўйи) федерал университети филиали ижрочи директори
Асатов Тимур Нурмухаматович
Жиззах шаҳридаги Қозон (Волгабўйи) федерал университети филиали ишлар бошқармаси бошлиғи
Хайдаров Озоджон Асламкулович
Creative Associates International TPD Lead

Аннотация

Мазкур мақолада маълумотларнинг интеллектуал таҳлили ва катта ҳажмдаги маълумотларга дастлабки ишлов бериш масалалари ўрганилган. Маълумотларга дастлабки ишлов беришга қаратилган кўплаб илмий тадқиқот ишлари назарий таҳлил қилинган.

Маълумотларни дастлабки ишлов бериш асосий соҳалари ўрганилиб, уларни турлари бўйича гуруҳга бўлинган ва ҳар бирига нисбатан мавжуд муаммолар кўрсатиб берилган.

Калит сўзлар: CNN, дискретлаштириш, Spark, Mllib, номукаммал маълумотлар, баланссиз маълумотлар.

Аннотация

В данной статье рассматриваются вопросы интеллектуального анализа данных и предварительной обработки больших объемов данных. Теоретически проанализированы многие научные исследования, посвященные предварительной обработке данных. Изучены основные области предварительной обработки данных, они сгруппированы по типам и указаны существующие проблемы, связанные с каждым.

Ключевые слова: CNN, дискретизация, Spark, Mllib, несовершенные данные, несбалансированные данные.

Annotation

This paper discusses data mining and preprocessing of large amounts of data. It theoretically analyzes many scientific studies devoted to data preprocessing. It examines the main areas of data preprocessing, groups them by type, and identifies the existing problems associated with each.

Keywords: CNN, discretization, Spark, Mllib, imperfect data, imbalanced data.

Республикамиз маълумотларга ишлов беришнинг янги даври интелектуал таҳлиллаш ва улардан фойдаланиш тизимларини ривожлантириш борасида янги даврга қадам қўймоқда. Бу қадамларни электрон тўловлар, электрон дўкон, аҳоли мурожаатлари билан ишлашни рақамлаштириш ва электрон ҳукумат тизими кабиларда кўриш мумкин. Мазкур йўналишлар марказлашган ҳолда ечимлар топиш механизми ҳисобланиб, макро ёндашув сифатида қараш ўринли. Иқтисодиётда бўлгани сингари макро ва микро иқтисодиёт тушунчаларига ажратилгани сингари, республика миқиёсида рақамлаштиришни макро йўналишга таққосланса, ҳудудий қамровда қаралиши микро йўналишга таққосланса ўринли бўлади. Муаммолар қуйи қатламда икир-чикиригача ўрганилса, ечим мустаҳкам ва мослашувчан бўлади. Қай бир бўғинда ўзини яҳши кўрсатган ечим бошқа бўғинларда кичикроқ ҳаражат билан қўллаб кўрилиши имконияти бўлади. Бу эса ҳар томонлама қулай ёндашув бўлади.

Бугунги кунда ишлов бериши зарур бўлган маълумот массивлари катта ҳажмни ташкил этаётгани оддий ҳолат бўлиб қолмоқда. Инсон ўзининг турли йўналишдаги фаолияти давомида турли кўринишдаги ахборотларни истеъмол қилади, бу жараёнда турли аспектларда ва турли мақсадларда маълумотга ишлов беришни кўзатиш мумкин. Катта массивли маълумотлар, глобал тармоқ ахборот маконида ахборот жамланмаларидаги маълумотлар турли кўринишда бўлиши, унга ишлов беришда маълум бир муаммолар келтириб чиқаради.

Юқоридаги мулоҳазалардан келиб чиқган ҳолда ҳудудий кесимдаги масалаларни ечиш учун BI (Business Intelligence) тизимлари тамоилларига асосланган таҳлиллаш ёки интелектуал таҳлиллаш тизимини ишлаб чиқиш ижтимойи ва иқтисодий соҳалар учун аҳамиятли ҳисобланади. Бундай тизимнинг тўғри ва самарали ишлаши маълумотларни дастлабки ишлов бериш механизмининг тўғри ташкил этилишига асосий эътибор қаратилиш лозим.

Маълумотларни интелектуал таҳлиллаш тизимларини қуришнинг умумий ёндашувлари. Маълумотларга дастлабки ишлов беришга қаратилган кўплаб илмий тадқиқот ишлари олиб борилган бўлиб, назарий таҳлил асосида уларнинг бир қаторига тўҳталиб ўтилади.

Ҳозирда маълумотлар оқими катта тезлик билан катталашиши унга тез ва сифатли ишлов бериш механизмларини ишлаб чиқиш, мавжудларини такомиллаштириш, турли типлиликни ва мураккаб тузилмаликни бартараф этиш ёндашувларига эҳтиёж ошиб бормоқда.

Маълумотларни интелектуал таҳлил қилишда ўзгарувчан муҳит бўлганлиги учун уларга ишлов беришда муаян бир усул билан ечим топишга харакат қилиш нотўғри ёндашув ҳисобланади. Бир ҳил усулнинг қўлланилиши натижалардаги ҳатоликларни катталашишига олиб келиши кузатилади. Бу эса катта массивли маълумотларни интелектуал таҳлил қилишда юқори самарадорликка эришилишини таъминлаш учун маълумотларга дастлабки ишлов бериш усул ва воисталарини такомиллаштиришга эҳтиёж туғдиради. Тадқиқот жараёнида бир қатор ишларни кўриб урғу берилиши зарур бўлган жабҳаларни белгилаб олишга харакат қилинди. Шундай қилиб, маълумотларга дастлабки ишлов бериш маълумотлар таҳлилидаги энг муҳим босқичлардан бири сифатида қаралади [1].

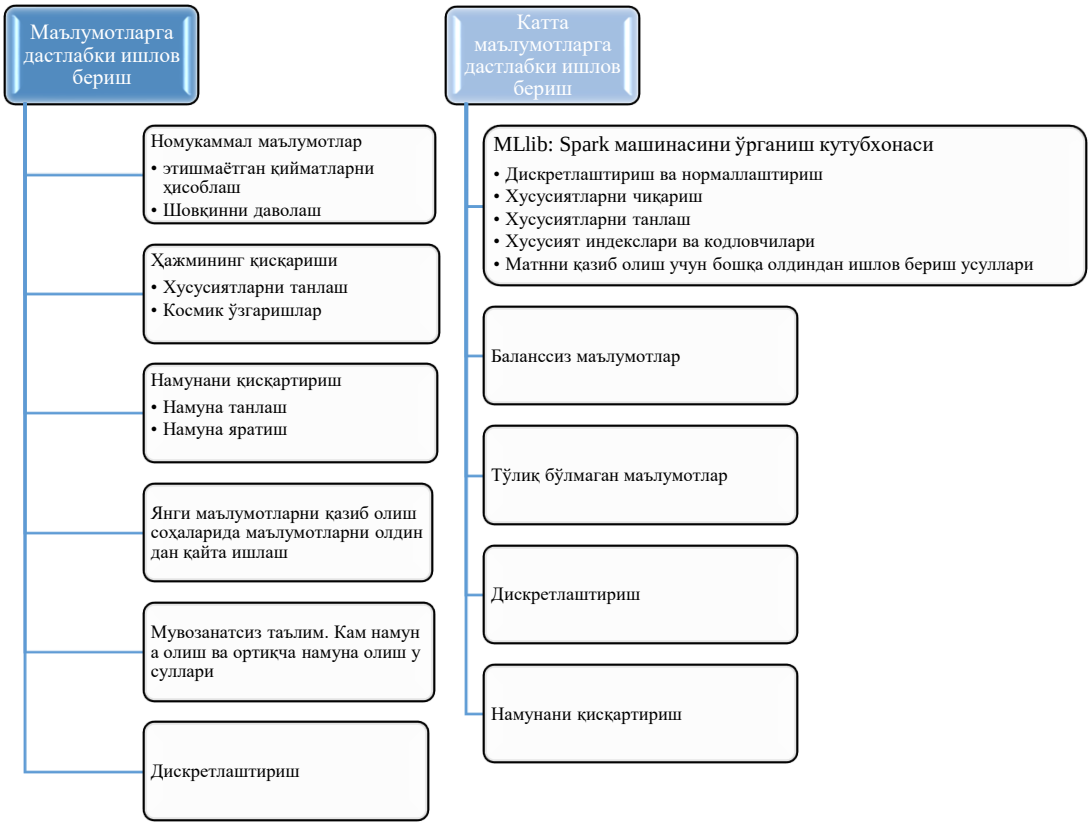
Маълумотларга дастлабки ишлов бериш фундаментал ёндашув бўлиб, нафақат ахборот технологиялари сохасида балки турлича, жумладан, қурилиш, энергетика, педагогика, юриспруденция ва транспорт сохаларида файдаланиш мумкин. Турк олимлари томонидан олиб борилган тадқиқотларда маълумотларга дастлабки ишлов беришда энергия истеъмоли маълумотларини кластерлаш ёндашуви қўлланилган [2].

Бошқарув қарорларини қабул қилишга кўмаклашувчи қарор муқобиллари самарадорлигини оширишда энг катта масалалар сирасига тоза маълумотларни аниқлаш ва уни тузатиш тадбирини амалга ошириш муҳим аҳамият касб этади [3,4,5]. Сўнги йилларда маълумотларни тозалашнинг янгидан-янги кўринишларини ривожлантириш истиқболлари очилмоқда, буларга содда бўлган статистик усулларнинг абстракция, интерфейс, масштаблаш учун янгиланаётган ёндашувларида намоён бўлган [6,7,8,9]. Бир қатор ишларда [10,11,12] статистик усуллардан фойдаланган ҳолда дастлабки шакллантирувчи маълумотларни тозалашни амалга оширишда машинали ўқитиш ёндашувларидан фойдаланиш қараб ўтилган бўлиб, ушбу тадқиқотлар истиқболли ва ўзига ҳос жиҳатларини белгилаб олиш ва ривожлантириш йўналишларини аниқлашда қўл келиши кўрсатилган.

Ҳар қандай ҳолатда статистика асосида тўпланган маълумотлар таркиби ва мазмунини, унинг шаклан тўғри дея олмаймиз. Шунинг учун дастлабки ишлов бериш натижа самарадорлигини оширишга йўналтирилган муҳитни ҳосил қилиш талаб этилади [13,14]. Маълумотларни таҳлилини (айниқса, статистикада) амалга оширишда уни тозалаш “ҳусусият” деб аталувчи кузатишлар асосида тўпланган маълумотлар устунни ҳисобга олишни назарда тутиш ўринли ҳисобланади [15]. Бу ерда ҳусусиятлар баёни чекловларни ҳисобга олиш ва уларга риоя қилинишни таъминлаш, анъанавий ёндашувлардаги миқдорий ўзгаришларни ҳисобга олинишидан бироз бошқача мазмун касб этсада, сезиларли ўзгариш бериши аниқланган. Ушбу ёндашув [13,15] ишларда ўз аксини топган. Таҳлиллашда конвалюцион нейрон тармоқлар (CNN) ёрдамида белгилар фазосини сифатли кўринишга келтириш масаласи [16] ишда қараб ўтилган бўлиб, таҳлиллагични CNN ёрдамида кўчайтирган.

Маълумотлар тузилмасидаги тезкор ўзгаришлар, сақланаётган маълумотларни тўғри баҳолаш ва улардан фойдали маълумотларни ажратиб олиш самарали дастлабки маълумотлар тўпламини ҳосил қилиш учун замин бўлади. Бунда, айниқса, тўлиқ бўлмаган маълумотларни тулдириш, ортиқчаларидан қутулиш муҳим вазифа саналади [17,18]. [19] ишда эса бу жараёнларни амалга ошириш турли кўринишдаги кластеризация ёндашувларини қўллаган ҳолда амалга оширилган. Шу билан бирга маълумотларга дастлабки ишлов беришда нотўғри, тўлиқ бўлмаган ва ноаниқликлар бўлган маълумотларга ишлов беришда норавшан ёндашувларнинг қўлланилиши ва самарали натижадорликка эришиш мумкинлиги кўрсатилган [20].

Маълумотлар массивида (бу ерда катта массивли маълумот ҳолати) машинали ўқитиш орқали тушуниш ва тушунчаларни ҳосил қилиш потенциал метамаълумотларни шакллантириш учун сифатли маълумотга эга бўлиш ёндашувлари [21,22,23,23] ишларда қараб ўтилган.



1-расм. Маълумотларга дастлабки ишлов бериш усуллари тавсифи

Яна кўришимиз мумкинки, биринчи 10 таликда ривожланган давлатлар ўрин олган бўлиб, ривожланиаётган мамлакатлар, жумладан, Ўзбекистонда ҳам бу соҳада изланишлар етарли эмаслигини тасдиқлайди.

Маълумотларни дастлабки ишлов бериш асосий соҳалари ўрганилиб, уларни турлари бўйича гуруҳга бўлинган ва ҳар бирига нисбатан мавжуд муаммолар кўрсатиб берилган. Дастлаб, етишмаётган қийматлар ва шовқинли маълумотлар киритилган номукаммал маълумотлар билан ишлаш учун дастлабки ишлов бериш усулларини кўриб чиқилган. Кейинги ўринларда, маълумотларни қисқартириш учун дастлабки ишлов бериш ёндашувлари тақдим этилган, унда хусусиятларни танлаш ва маконни ўзгартириш кўрсатиб ўтилган. Кейинги бўлимда белгиларни қисқартириш алгоритмлари, жумладан белгиларни танлаш ва прототип яратиш кўриб чиқилган. Охирги учта бўлим мос равишда дискретлаш, номутаносиб масалалар учун қайта белгиларни шакллантириш ва маълумотларга ишлов беришнинг янги йўналишларига бағишланган (1-расм).

Олиб борилаётган тадқиқотлар салмоғи баҳоси. Турли соҳаларда маълумотларга дастлабки ишлов беришни ўрганишлар давомида барча давлатлар олимларининг чоп этган мақолалари ва илмий изланишлари кўриб чиқилди. Жумладан www.webofscience.com платформаси электрон журнали орқали мавзуга оид мақолалар таҳлил қилинганда, 1011 та мақола борлиги аниқланди (1-жадвал). Мавжуд йилмий мақолалар 88 та давлатнинг олимлари томонидан ўрганилганлигини кўрсатди.

1-жадвал. Дунё давлатларининг webofscience платформаси базасида предмет соҳа бўйича нашр этиш ҳолати.

1	Хитой Халқ Республикаси - 329		45	Чили - 4
2	АҚШ - 134		46	Ирландия - 4
3	Жанубий Корея - 84		47	Исроил - 4
4	Саудия Арабистони - 76		48	Литва - 4
5	Ҳиндистон - 59		49	Таиланд - 4
6	Германия - 53		50	Украина - 4
7	Англия - 48		51	Болгария - 3

8	Италия - 47		52	Хорватия - 3
9	Испания - 40		53	Гана - 3
10	Миср - 36		54	Венгрия - 3
11	Покистон - 33		55	Иордания - 3
12	Канада - 28		56	Қозоғистон - 3
13	Туркия - 28		57	Қувайт - 3
14	Япония - 27		58	Нигерия - 3
15	Эрон - 26		59	Шотландия - 3
16	Малайзия - 25		60	Словения - 3
17	Тайван - 25		61	Уелс - 3
18	Норвегия - 18		62	Жазоир - 2
19	Франция - 17		63	Ангола - 2
20	Нидерландия - 17		64	Аргентина - 2
21	Полша - 17		65	Бруней - 2
22	Финляндия - 16		66	Эстония - 2
23	Австралия - 15		67	Латвия - 2
24	Индонезия - 13		68	Непал - 2
25	Швеция - 12		69	Шимолий Ирландия - 2
26	Чехия - 11		70	Филиппин - 2
27	Греция - 11		71	Қатар - 2
28	Мексика - 11		72	Тунис - 2
29	Португалия - 11		73	Ветнам - 2
30	Швейцария - 11		74	Камерун - 1
31	Бангладеш - 10		75	Фил суяги - 1
32	Белгия - 10		76	Куба - 1
33	Ироқ - 10		77	Доминикан вакили - 1
34	Бразилия - 9		78	эквадор - 1
35	Россия - 9		79	Сальвадор - 1
36	Дания - 7		80	Эфиопия - 1
37	Бирлашган Араб Амирликлари - 7		81	Исландия - 1
38	Австрия - 6		82	Кения - 1
39	Колумбия - 6		83	Люксембург - 1
40	Марокаш - 6		84	Мадагаскар - 1
41	Руминия - 6		85	Мўғулистон - 1
42	Словакия - 6		86	Уммон - 1
43	Жанубий Африка Республикаси - 6		87	Ўзбекистон - 1
44	Сербия - 5		88	Зимбабве - 1

Хулоса. Маълумотларнинг экспоненциал ошиши, турларининг кўпайиши унга ишлов беришга қўйиладиган талабларни ўзгартиришга, бир нечта усуллардан бир вақтнинг ўзида фойдаланишни йўлга қўйиш зарурати бўйича тадқиқотларни кучайтириш лозимлигини кўрсатди. Маълумотларга дастлабки ишлов бериш сохасида олиб борилаётган илмий тадқиқот ишлари асосан ахборот технология, бизнес, техника, медицина йўналишлари ва ижтимойи соҳада кенг қамраб олинмоқда. Амалий жиҳатдан олиб борилган тадқиқотларнинг аксарияти иқтисодий кўрсаткичлар асосида қисқа ва ўрта муддатли башоратлаш, ҳолатларни баҳолаш орқали самарали қарор муқобиллари банкини шакллантириш каби мураккаб масалаларни ҳал этиш, янги механизмларни ўйлаб топиш, жараёнлар мураккаблигига қараб мослашувчан ишлов бериш муҳитини ҳосил қилишга талабни оширмоқда. Ўрганиш жараёнида республикамизда тадқиқотларнинг молия, иқтисодиёт, аудит ва бухгалтерия каби ҳаётий зарур жабҳаларида олиб борилган илмий изланишлар етарлича ривожланмаганлиги кузатилди. Республикани иқтисодий потенциалини кучайтиришда кенг қўлланилиши мумкин бўлган тезкор таҳлиллаш

тизимлари маълумотларга дастлабки ишлов бериш усул ва воситаларини иқтисодий ва молиявий таҳлилларига кенг қўлланилиши мақсадга мувофиқ. Бунда асосий масалалардан дастлабки маълумотларни тўғри шакллантириш билан бир қаторда улардаги турли ҳалақитлар, етишмовчиликлар, ортиқчаликлар каби навбатдаги босқич ишлов бериш жараёнларида самарали ечим олишга тўсқинлик қилувчи омилларни бартараф этиш механизмини ишлаб чиқиш, ишлов бериш тизимлари таркибида қўллаш ресурслардан оқилона фойдаланиш, сарф-ҳаражатларни камайтиришга эришишнинг энг мақбул йўли ҳисобланади. Ўтказилган тизимли таҳлил тадқиқотларидан ҳозирда мавжуд моно воситали механизмлардан кўра кўп вариантли воситаларни ишлаб чиқиш мақсадга мувофиқлигини кўрсатди.

Фойдаланилган адабиётлар рўйхати:

1. Obinwanne, Tobechi, et al. "Data Preprocessing in Supply Chain Management Analytics-A Review of Methods, the Operations They Fulfill, and the Tasks They Accomplish. Data Preprocessing in Supply Chain Management Analytics." *Proceedings of the 2023 6th International Conference on Computers in Management and Business*. 2023.
2. Yıldız M. Ş., Doğanşahin K., Kekezoğlu B. Data Preprocessing in Electrical Energy Consumption Profile Clustering Studies.
3. X.Chu, I.F.Ilyas, S.Krishnan, J.Wang Data Cleaning: Overview and Emerging Challenges//SIGMOD'16, June 26-July 01, 2016, San Francisco, CA, USA ©c 2016 ACM. ISBN 978-1-4503-3531-7/16/06. . . DOI: http://dx.doi.org/10.1145/2882903.2912574
4. M. Bergman, T. Milo, S. Novgorodov, and W. C. Tan. Query-oriented data cleaning with oracles. In SIGMOD, 2015.
5. A. Chalamalla, I. F. Ilyas, M. Ouzzani, and P. Papotti. Descriptive and prescriptive data cleaning. In SIGMOD, pages 445–456, 2014.
6. Z. Khayyat, I. F. Ilyas, A. Jindal, S. Madden, M. Ouzzani, P. Papotti, J. Quiané-Ruiz, N. Tang, and S. Yin. Bigdansing: A system for big data cleansing. pages 1215–1230, 2015.
7. C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. L. Roth. Preserving statistical validity in adaptive data analysis. In STOC, pages 117–126, 2015.
8. S. Krishnan, J. Wang, E. Wu, M. J. Franklin, and K. Goldberg. Activeclean: Interactive data cleaning while learning convex loss models. In Arxiv: http://arxiv.org/pdf/1601.03797.pdf, 2015
9. H. Xiao, B. Biggio, G. Brown, G. Fumera, C. Eckert, and F. Roli. Is feature selection secure against training data poisoning? In ICML, 2015.
10. A. Heise, G. Kasneci, and F. Naumann. Estimating the number and sizes of fuzzy-duplicate clusters. In CIKM Conference, 2014.
11. I. F. Ilyas and X. Chu. Trends in cleaning relational data: Consistency and deduplication. Foundations and Trends in Databases, 5(4):281–393, 2015.
12. M. Volkovs, F. Chiang, J. Szlichta, and R. J. Miller. Continuous data cleaning. In ICDE, pages 244–255, 2014.
13. Chicco D, Oneto L, Tavazzi E (2022) Eleven quick tips for data cleaning and feature engineering. PLoS Comput Biol 18(12): e1010718. https://doi.org/10.1371/journal.pcbi.1010718
14. Baillie M, le Cessie S, Schmidt CO, Lusa L, Huebner M. Ten simple rules for initial data analysis. PLoS Comput Biol. 2022; 18(2):e1009819. https://doi.org/10.1371/journal.pcbi.1009819 PMID: 35202399
15. Kilkenny MF, Robinson KM. Data quality: "Garbage in-garbage out". Health Inf Manag J. 2018; 47 (3):103–5. Ilyas IF, Chu X. Data cleaning. Morgan & Claypool; 2019.

16. Bai Y, Mei J, Yuille AL, Xie C. Are transformers more robust than CNNs? Adv Neural Inf Process Syst. 2021; 34:26831–43.

17. V.ÇETİN, O.YILDIZ A comprehensive review on data preprocessing techniques in data analysis// Pamukkale University Journal of Engineering Sciences , 28(2), 299-312, 2022

18. Ramírez-Gallego S, Krawczyk B, García S, Wozniak M, Herrera F. "A survey on data preprocessing for data stream mining: Current status and future directions". Neurocomputing, 239, 39-57, 2017

19. Shanthini A, Vinodhini G, Chandrasekaran RM. "A taxonomy on impact of label noise and feature noise using machine learning techniques". Soft Computing, 23, 8597-8607, 2019

20. Nwadiugwu M. "Gene-Based clustering algorithms: Comparison between denclue, fuzzy-C, and BIRCH". Bioinformatics and Biology Insights, 14, 1-6, 2020.

21. Kiebler, Lorenz; Moroff, Nikolas Ulrich; Jacobsen, Jens Jakob (2022) : Preliminary analysis on data quality for ML applications, In: Kersten, Wolfgang Jahn, Carlos Blecker, Thorsten Ringle, Christian M. (Ed.): Changing Tides: The New Role of Resilience and Sustainability in Logistics and Supply Chain Management – Innovative Approaches for the Shift to a New Era. Proceedings of the Hamburg International Conference of Logistics (HICL), Vol. 33, ISBN 978-3-756541-95-9, epubli GmbH, Berlin, pp. 207-236, https://doi.org/10.15480/882.4693

22. Alsheibani, S., Cheung, Y. and Messom, C., 2019. Factors Inhibiting the Adoption of Artificial Intelligence at organizational-level: A Preliminary Investigation. In: M. Santana, and R. Montealegre. AMCIS 2019 Proceedings. Americas Conference on Information Systems 2019. Cancun, Mexico, 15/08/19

23. Brownlee, J., 2020. Data preparation for machine learning: data cleaning, feature selection, and data transforms in Python: Machine Learning Mastery.

24. Khaydarov, V., Heinze, S., Graube, M., Knupfer, A., Knespel, M., Merkelbach, S. and Urbas, L., 2020. From stirring to mixing: artificial intelligence in the process industry. In: 2020 25th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA). Vienna, Austria, 9/8/2020 - 9/11/2020. Piscataway, NJ: IEEE, pp. 967–974