Raport Faza I (UTCN)

Sisteme de monitorizare a resurselor de apa

(Achiziția, colectarea, preprocesarea datelor si detecția anomaliilor)

Cuprins

1. Apa ca resursa primordiala

Este de necontestat faptul ca apa reprezintă unul din elementele esențiale necesare pentru menținerea oricărei forme de viată. Într-o societate civilizată administrarea resurselor și a consumurilor de apă este o responsabilitate majoră care influențează atât prezentul cât mai ales viitorul societății. Conform Adunării Generale a Organizației Națiunilor Unite, în iulie 2010 s-a stabilit că accesul la apă suficientă pentru uz personal și domestic este un drept al fiecărui om. Acest drept presupune printre altele ca fiecare om să aibă acces zilnic la un volum de apă de 50 până la 100 de litri [1].

Foarte mult timp, ca și în cazul altor resurse, s-a considerat că resursele de apă sunt nelimitate și există un proces natural de regenerare a acesteia. Dar o serie de studii [2] au arătat ca exploatarea neraţională a acestei resurse duce la degradarea acesteia și procesul de regenerare are loc într-o măsură mult mai mică decât consumul.

In foarte multe regiuni de pe glob lipsa apei sau reducerea drastică a precipitațiilor (cauzate mai ales de schimbările climatice) a dus la reducerea potențialului agricol și, implicit, la fenomene de sărăcire a populației. Organizația Națiunilor Unite estimează că 90% din dezastrele naturale sunt strâns legate de apă, inclusiv inundațiile și seceta [1]. Fenomenul, inițial considerat de importanță locală sau regională, devine o problema globală, datorită fenomenelor geopolitice pe care le generează: foamete/sărăcie, conflicte armate, migrări masive de populații etc.

Nu întâmplător o serie de organizații si agenții internaționale încearcă sa stopeze procesul de degradare a resurselor de apă prin dezvoltarea unor strategii pe termen mediu și lung pentru asigurarea unui echilibru între consumul și procesul de regenerare a resurselor de apă.

Uniunea Europeană prin Directiva cadru privitoare la apă (EU Water Framework Directive, https://ec.europa.eu/environment/water/fitness_check_of_the_eu_water_legislation/index_en.htm) stabilește pentru țările membre regulile generale privind modul de tratare unitară a diverselor aspecte legate de managementul diferitelor surse de apă. Există documente directoare pentru administrarea bazinelor hidrografice și controlul inundațiilor, pentru administrarea apei de consum pentru populație, administrarea deșeurilor de apă canalizare), managementul apei utilizate pentru agricultura, controlul parametrilor de calitate ai apei și administrarea sistemului hidrografic marin. Prin așa-numitul pachet “Fittness check” elaborat în 2019 se urmărește acordarea legislațiilor naționale din țările membre cu privire la regulile generale de administrare și control al calității apei. Prin acțiunea denumita “Inception Impact Assessment”, derulată în perioada 23 octombrie - 20 noiembrie 2020 se urmărește evaluarea stadiului în care directiva europeană cu privire la managementul apei se regăsește în legislația și în strategia guvernamentală a țărilor membre.

Prin natura lor, resursele de apă nu pot fi administrate doar la nivel regional sau național. Bazinele hidrografice nu țin cont de granițe sau de interese sau influențe geo-politice. Cursurile unor fluvii (vezi Europa, Asia sau Africa) traversează mai multe țări și politica de administrare a unei regiuni influențează accesul la această resursă pentru celelalte țări riverane. Politica rațională de protejare și regenerare a resurselor de apă promovate de anumite țări cu economie avansată nu este suficientă, dacă nu se acordă sprijin financiar și tehnologic pentru unele regiuni mai puțin dezvoltate. Alte ori o politică agresivă de exploatare intensivă a cursurilor de apă într-o țară (vezi China sau Israel) afectează necesitățile de apă și implicit situația socioeconomică a unor țări vecine.

Chiar si in cadrul aceleiași tari, precum Romania, mijloacele de administrare a resurselor de apa si de monitorizare a consumurilor pot sa varieze in funcție de regiune sau de contextual urban sau rural. Deși tot mai multe conglomerate urbane se raliază la ideea de orașe inteligente in care managementul rețelelor de apa joaca un rol esențial, exista localități in care nu exista apa curenta (“la robinet”) a cărei calitate sa fie controlata si garantata. In foarte multe localități sistemul de colectare a apelor uzate nu exista sau, daca exista, apa reziduala nu este tratata adecvat.

Figura 1.1 prezinta situația gradului de tratare a apelor reziduale in tarile europene.

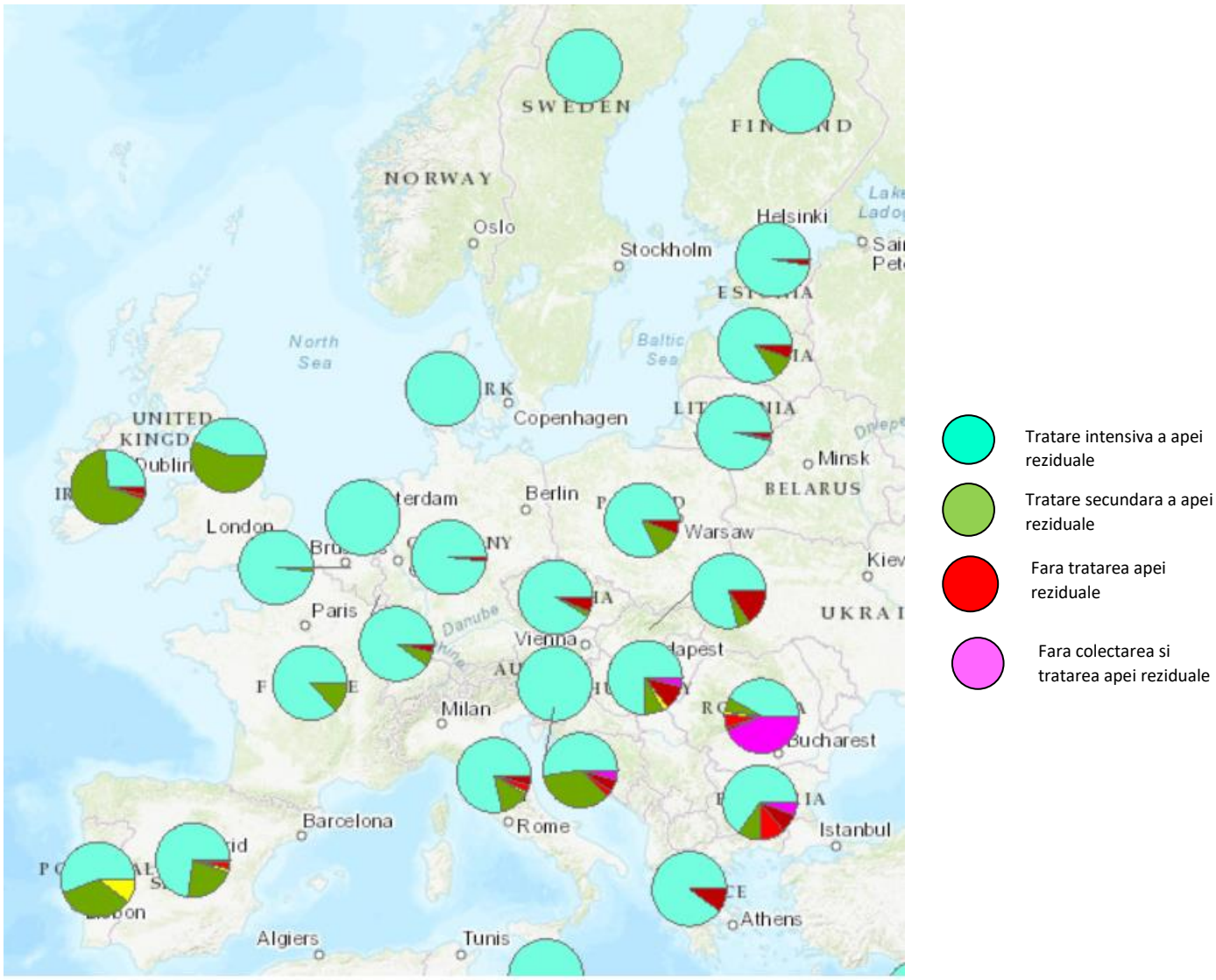

Figura 1.1 Situația gradului de tratare a apelor reziduale in tarile europene

Accesul la apa curenta de calitate ar trebui sa fie un minim in Europa secolului 21 si mijloacele de distribuire si regenerare eficienta a acestei resurse ar trebui sa fie prioritatea oricărei guvernări. Tehnologia pentru a face acest lucru exista si soluțiile pilot de administrare inteligenta a resurselor de apa implementate in cadrul unor proiecte de tip „oraș inteligent” ar trebui generalizate.

In acest context **scopul proiectului de fata este acela de a oferi soluții avansate de monitorizare si administrare a resurselor de apa**, soluții care implica combinarea unor tehnici avansate din domeniul sistemelor de calcul si al comunicațiilor. In mod concret, in cadrul proiectului se vor utiliza tehnici avansate de colectare automata a datelor (prin tehnologii IoT, smart sensing si rețele senzoriale fără fir), de procesare si catalogare inteligenta a datelor (prin algoritmi de inteligenta artificial) si de generare a unor informații sintetice, necesare ca suport pentru luarea de decizii optimale.

1.1. Politici/strategii globale, regionale si naționale de prezervare si utilizare eficienta a resurselor de apa

La nivel global unul dintre cele 6 puncte țintă ale ONU este ca până în anul 2030 să se implementeze sisteme de management integrat al resurselor de apa la toate nivelele inclusiv prin cooperare transfrontalieră [4]. Un recent raport al Organizației Națiunilor Unite ("UN World Water Development Report 2020 'Water and Climate Change'") sintetizează principalele problem actuale cu care se confrunta tarile lumii cu privire la administrarea resurselor de apa in contextul schimbărilor climatice. Printre aspectele tratate se pot enumera: corelația dintre schimbările climatic si modificările din sistemele hidrografice regionale, importanta administrării apelor reziduale, influenta calității apei asupra sănătății populației, si masuri de combatere a degradării surselor de apa. Raportul constata prezinta o serie de date statistice privind satisfacerea nevoilor de apa in diferite regiuni geografice; astfel statisticile arata ca in prezent, 2,2 miliarde de oameni nu au acces la apă potabilă gestionată în condiții de siguranță, iar 4,2 miliarde, sau 55% din populația lumii, nu dispun de instalații sanitare gestionate în condiții de siguranță.

Din perspectiva Băncii Mondiale, „Apa atinge fiecare aspect al dezvoltării și se leagă de aproape fiecare obiectiv de dezvoltare durabilă. (Aceasta resursa) Determina creșterea economică, susține ecosisteme sănătoase și este esențiala și fundamentala pentru viața însăși.... Serviciile de apă, igienizare și igienă (WASH) gestionate în siguranță sunt o parte esențială a prevenirii și protejării sănătății umane în timpul focarelor de boli infecțioase, inclusiv a pandemiei actuale COVID-19."

La nivelul comunității europene există încă de la nivelul anului 2000 o directiva cadru "EU Water Framework Directive" pentru o politica comună a statelor membre în domeniul apei [3]. Aceasta are, printre altele, declarate ca obiective următoarele două aspecte cheie: fiecare țară membră să monitorizeze starea apei din fiecare bazin și să analizeze caracteristicile fiecărui bazin hidrografic, inclusiv impactul activității umane și să realizeze o evaluare economică a utilizării apei. Directiva stabilește pași concreți de implementare a acestei politici pana in 2030, in vederea realizării sistemului de management si de monitorizare unitara a resurselor de apa din Europa.

Deși conform infograficului Comisiei Europene pentru marea majoritate a statelor europene parametrii microbiologici și chimici ai apei potabile sunt respectați din perspectiva nivelelor maxime admisibile legale în proporție de 99% până la 100%, totuși la nivel european peste 40% din cursurile de apa sunt expuse presiunilor hidromorfologice [4].

La nivel național, politicile legate de administrarea resurselor si consumurilor de apa se integrează in "Strategia naționala de dezvoltare durabila a României 2030". In acest act director, un obiectiv special (obiectivul nr. 6 "Apa curata si sanitație") este "asigurarea disponibilității si gestionarii durabile a apei si sanitație pentru toți". Printre aspectele evidențiate se numără:

- Creșterea calității vieții prin dezvoltarea infrastructurii de apa si canalizare
- Alinierea României la cerințele si standardele UE gestionarea apei potabile, a apelor uzate si a deșeurilor.
- Creșterea eficientei utilizării apei in toate sectoarele

Cadrul legislative pentru implementarea acestor obiective este dat de Legea nr. 458/2002 modificata si actualizata in 2017 ce vizează mai ales aspectele privind respectarea parametrilor minimi de calitate pentru apa potabila. Aceasta lege este un rezultat al Protocolului privind apa si sănătatea, protocol semnat de România cu Organizația Mondiala a Sănătății. Conform [5] „calitatea apelor din România este urmărită conform structurii și principiilor metodologice ale Sistemului de Monitoring Integrat al Apelor din România (S.M.I.A.R.). Astfel se realizează o monitorizare de supraveghere având rolul de a evalua starea tuturor corpurilor de apă din cadrul bazinelor hidrografice și un monitoring operațional (integrat monitorizării de supraveghere) pentru corpurile de apă ce au riscul sa nu îndeplinească obiectivele de protecție a apelor.

1.2. Necesitatea monitorizării si administrării resurselor si a consumurilor de apa

Monitorizarea resurselor și a consumului de apă presupune efectuarea de măsurători cu privire la cantitățile de apă produse și consumate, a parametrilor fizici și chimici ai apei urmată de analiza, procesarea și raportarea rezultatelor acestor măsurători către părțile interesate. Monitorizarea este un proces de lungă durată care conduce la acumularea unor cantități mari de date și realizarea unui istoric în timp a tuturor parametrilor monitorizați care permit înțelegerea unor fenomene fizice, a unor procese ciclice și aduce feedback în urma întreprinderii de măsuri de modificare a ecosistemului monitorizat [6].

Monitorizarea resurselor de apă este necesară pentru administrarea acestora, protecția mediului, identificarea calității apei, a eventualelor surse de poluare și nu în ultimul rând pentru educarea întregii comunități care este deservită de sursele de apă. Astfel, se poate îmbunătății considerabil atitudinea întregii comunități cu privire la consumul apei și cu privire la calitatea acesteia. Nu în ultimul rând monitorizarea conduce în cele din urmă și la o reducere a costurilor cu resursele de apă pentru toate părțile implicate în producerea, transportul și consumul apei.

Necesitatea si actualitatea masurilor de monitorizare si management global al apei decurg dintr-o serie de date statistice si constatări ale unor organisme internaționale de specializate. De exemplu conform datelor publicate de Națiunile Unite [7]:

- La nivel global, peste 80% din apele uzate sunt eliberate în mediu fără o tratare adecvată

- Costurile gestionării apelor uzate sunt mult sub beneficiile aduse pentru sănătatea umană, dezvoltarea economică și durabilitatea mediului; lipsa apei potabile de calitate duce la răspândirea de boli specifice

- aproximativ 1,8 miliarde de oameni folosesc o sursă de apă potabilă contaminată cu fecale, punându-le în pericol de a contracta holera, dizenteria, tifoida și poliomielita

- cele mai multe probleme legate de calitatea apei sunt cauzate de agricultura intensivă, producția industrială, exploatarea minieră și scurgerile urbane netratate și apele uzate.

- Schimbările climatice din ultimele decenii afectează negativ resursele de apa; in diferite regiuni apar fenomene de seceta si de degradare a terenurilor agricole

Aceste constatări sunt in parte valabile si pentru Romania, mai ales pentru mediul rural. Pentru mediul urban problema principala este cea a tratării si reciclării apelor uzate, menajere si industriale. Din acest motiv **consideram ca tehnologiile avansate din domeniul tehnicii de calcul, al comunicațiilor si al sistemelor automate permit astăzi realizarea unor sisteme distribuite de monitorizare si administrare eficienta a resurselor de apa**. **Prin tehnici avansate de colectare a datelor (ex: prin tehnologii IoT, rețele senzoriale fără fir, edge computing), de procesare inteligenta a informațiilor si de modelare, simulare si previzionare a fenomenelor specifice pot fi implementate strategii optime de administrare a resurselor de apa. Prin tehnici de detecție automata a anomaliilor pot fi identificate si soluționate in timp util situații sau fenomene critice (ex. poluări accidentale, inundații).**

2. Sisteme de monitorizare a resurselor de apa

Sistemele de monitorizare a rețelelor hidrografice și de distribuție a apei prezintă multiple provocări din punct de vedere structural. Pe de-o parte se efectuează monitorizarea la nivelul surselor de apă și a punctelor de acumulare a acesteia, iar pe de altă parte trebuie monitorizată o întreagă rețea de distribuție și de consum a apei.

Prima componentă presupune existența unor stații automatizate sau semi-automatizate de măsurare a resurselor de apă din diferite surse cum sunt barajele sau digurile, prizele care colectează apa din surse de suprafață, cum sunt râurile, lacurile și lacurile de acumulare și nu în ultimul rând măsurarea debitelor la derivațiile cursurilor de apă.

Cea de-a doua componentă, partea de distribuție și de consum presupune existența unei rețele complexe de țevi de diferite circumferințe și din diferite materiale, cel mai adesea îngropate și inaccesibile echipamentelor de măsurare. În cele mai multe cazuri, măsurătorile cantitative și calitative se obțin cel mai ușor la punctele de consum final, dar și aceste măsurători pot fi un bun indicator pentru detecția problemelor în alte puncte inaccesibile alea rețelei. Dintre aceste probleme merită a fi menționate cel puțin cele legate de deteriorarea țevilor pe anumite tronsoane care duce la pierderi de apă sau în cazul țevilor metalice mai vechi la deteriorarea parametrilor de calitate a apei. Tot cu ajutorul datelor obținute la punctele de control se pot identifica în mod preemptiv segmente considerate mai susceptibile cedării structurale sau puncte „problemă” de-a lungul rețelei.

La nivel național colectarea datelor din cadrul rețelelor hidrografice se face prin intermediul a 4 tipuri de stații de măsurare: hidrologice, evapometrice, de analiză a calității apei și de analiza a apei rezultate din precipitațiile atmosferice lichide sau solide. Începând cu anul 2013 in cadrul proiectului WATMAN [8], la nivel național s-au pus în funcțiune mai multe zeci de stații de monitorizare complet automate. De asemenea, la nivelul companiilor locale de distribuție de apa exista sisteme de monitorizare a rețelei de distribuție (ex: presiune apa, consum apa, detectarea pierderilor in rețea) si a procesului de curățare a apelor uzate (ex: Compania de apa Someș SA derulează proiecte in acest sens https://www.casomes.ro/).

2.1. Caracteristici generale ale unui sistem de monitorizare a resurselor de apa

Așa cum este prezentat in [62], un sistem inteligent de monitorizare a resurselor de apa cuprinde următoarele componente de baza:

- **Sursele de date**: acestea pot fi dispozitive individuale care realizează măsurători (senzori), rețele de senzori conectate la Internet prin dispozitive de tip gateway, sau chiar surse externe de date generate de alte sisteme.

- **Canale de comunicație cu dispozitivele**: asigura transmisia bidirecționala a datelor generate de senzori înspre serverul central, precum si transmisia comenzilor in direcția opusa, folosind protocoale de comunicație fără fir. Pentru a realiza acest lucru este nevoie de dispozitive gateway multi-protocol care joaca rolul de noduri intermediare.
- **Componenta de administrare a dispozitivelor**: este responsabilă cu păstrarea de informații despre starea surselor de date, de control acestora de la distantă si de supraveghere a dispozitivelor gateway.
- **Componenta de administrare a datelor**: primește, decodifica, decriptează si stochează datele trimise de sursele de date.
- **Componenta de procesare si analiza a datelor**: este responsabila pentru procesarea, analiza datelor si de trimiterea rezultatelor la componenta care se ocupa de stocare datelor. Aceasta componenta poate include metode statistice de analiza a datelor, metode de predicție si metode de învățare automata.
- **Componenta de interacțiune cu utilizatorul**: Utilizatorii unui astfel de sistem pot fi operatori din cadrul companiilor care gestionează resursele de apa sau consumatori ai resurselor de apa. Fiecare grup de utilizatori are un set de cerințe distinct si perspective diferite in utilizarea sistemului. Companiile au ca obiectiv monitorizarea si analiza resurselor de apa consumate, starea dispozitivelor si controlul acestora de la distanta. Consumatorii au ca obiectiv urmărirea datelor proprii, compararea cu medii ale consumatorilor din aceeași zona si cu același profil de utilizare si, eventual, urmărirea stării si controlul dispozitivelor instalate pe proprietatea lor.

2.2. Senzori si dispozitive de măsurare a mărimilor fizice asociate rețelelor de apa

Pentru monitorizarea si administrarea rețelelor de apa este necesara urmărirea mai multor mărimi fizice specifice, precum:

- debit de apa
- nivelul apei
- cantitate de apa consumata/deversata
- nivel/cantitate de apa in diferite incinte/rezervoare
- presiune statica si dinamica; presiune diferențiala
- compoziție chimica si fizica a apei: pH, concentrație de nitrați si nitriți, produși cu fosfor, metale grele, etc.
- temperatura apei, temperatura aerului exterior în punctele de măsurare
- nivelul de precipitații lichide sau solide

Pentru măsurarea integrității rețelelor de apa si identificarea unor defecțiuni se utilizează diverse fenomene fizice care pot fi evaluate/interpretate prin măsurători specifice: laser, unde acustice, unde radio, mărimi electrice (rezistenta, inductanța, capacitate), etc.

Măsurarea acestor parametric in regim continuu si in mod automat presupune utilizarea unor senzori specifici si a unei infrastructuri de colectare a datelor.

Alegerea tipului de senzor se va face in funcție de:

- Natura mărimii fizice măsurate
- Plaja de variație admisa a mărimii fizice măsurate

- Acuratețea admisa sau eroarea maxima admisibila
- Timpul de răspuns maxim al senzorului
- Caracterul continuu sau discontinuu al procesului de măsurare
- Condiții de protecție pentru dispozitivul de măsură (clase de protecție pentru praf, apa)
- Modul de transmitere a valorii măsurate

In funcție modul de transmitere a mărimii măsurate, respectiv a mărimii fizice de ieșire, senzorii se pot clasifica in:

- senzori analogici – a căror ieșire este o valoare de tensiune sau curent care variază într-o plaja predefinita (ex. 0-10V, -5 - +5V, 4-20mA); distant de transmisie a informației măsurate este limitata (10-50m), iar semnalul analogic este influențat de zgomotele electromagnetice din mediul in care se transmite informația; acești senzori pot fi utilizați in soluții locale de monitorizare si control dar nu se pretează pentru culegerea de date pe arii geografice extinse.
- senzori inteligenți - care au capabilitatea de a transmite informația culeasa pe un canal digital de comunicație; codificarea informației măsurate se poate realiza prin:
    - doua stări predefinite (ex. Închis/deschis, prezenta/lipsa presiune, etc.)
    - frecventa variabila a impulsurilor
    - dimensiune variabila a impulsurilor (ex. tehnica PWM)
    - valori sau mesaje transmise printr-un canal serial dedicat sau multiport (ex. RS232, RS485)
    - mesaje transmise într-o rețea de comunicație

Din perspectiva realizării unor sisteme de monitorizare care acoperă o arie geografica mare (de ordinal kilometrilor) si care au in component un număr mare de senzori (peste 100) se recomanda utilizarea senzorilor inteligenți care au capabilitatea de a comunica pe o rețea industrial de comunicație. Un senzor inteligent are de obicei un preț suplimentar in comparație cu un senzor analogic, însă acest cost suplimentar este compensat de cheltuielile mult mai mici necesare pentru realizarea infrastructurii de colectare a datelor. In plus, un senzor inteligent are funcționalități suplimentare care contribuie la creșterea acurateței de măsurare si asigura o fiabilitate sporita (prin tehnici de auto calibrare, filtrare a datelor, eliminarea erorilor de neliniaritate, etc.). De asemenea, datele pot fi stocate temporar, pot fi afișate local si pot fi transmise într-o forma împachetata.

O soluție de colectare a datelor bazata pe o rețea industriala de comunicării utilizează un singur mediu de comunicație (cu sau fără fir) pentru conectarea unui număr mai mare de senzori si prin protocolul de comunicație se poate verifica integritatea datelor transmise si a echipamentelor aflate la distanta. O rețea industriala de comunicații utilizează un protocol adaptat necesităților specific de colectare a datelor din cadrul unor sisteme complexe de automatizare. Structura mesajelor transmise (de dimensiuni reduse), periodicitatea traficului, regulile deterministe de acces la mediul de comunicație sunt doar câteva caracteristici ale rețelelor industriale, exploatabile in cazul sistemelor de monitorizare a instalațiilor de apa.

O soluție alternativa de colectare a datelor prin senzori/dispozitive inteligente este utilizarea rețelei Internet si a unor dispozitive care pot fi conectate direct pe o astfel de rețea. In acest caz vorbim de soluții de tip IoT (Internet of Things) sau IIoT(Industrial IoT), foarte populare in ultima perioada. Exista o serie de companii specializate in fabricația de senzori si dispozitive de măsură care oferă variante constructive

conectabile la Internet, fie in mod direct fie prin niște dispozitive de adaptare (gateway). Un dispozitiv de adaptare poate deservi mai mulți senzori conectați pe o rețea locala industriala.

Utilizarea Internetului in monitorizarea rețelelor de apa permite acoperirea unor arii geografice mari, asigura integrarea ușoara cu alte sisteme de monitorizare sau de control al serviciilor utilitare si datele colectate devin disponibile pentru o categorie mai larga de potențiali utilizatori.

2.2.1 Modele de senzori si dispozitive de măsură

In continuare se prezinta cu titlu informativ câteva tipuri de senzori utilizabili pentru monitorizarea rețelelor de apa.

A. Exemple de senzori si sisteme automate pentru măsurarea cantității de apa consumata

- Sistemul Itron-EquaScan – caracteristici:
  - Utilizează cate un dispozitiv radio montat pe fiecare contor de apa si o stație de citire portabila
  - Culegerea de la distanta a consumului de apa in sistemul „walk-by”
  - Transmisie radio 868,9 MHz
  - Descărcarea datelor culese într-un PC prin Bluetooth
  - Funcții de siguranța: detecția curgerii inverse, detecție tentativă de furt
- AquaWAN – caracteristici:
  - Sistem complet de monitorizare de la distanta a consumurilor de apa
  - Utilizează o soluție IoT cu un protocol de comunicație LoRaWAN
  - LoRa – protocol de comunicație radio, Long Range (acoperă distante de ordinul zecilor de km), Low power
  - Banda de frecventa 868MHz
  - Consum foarte mic in regim „stand-by”; bateria rezista pana la 10 ani
  - Posibilitatea de a închide circuitul de apa
  - Măsoară mai mulți parametri fizici: consum, presiune, debit, sau de calitate a apei
  - Furnizor: eta-2u

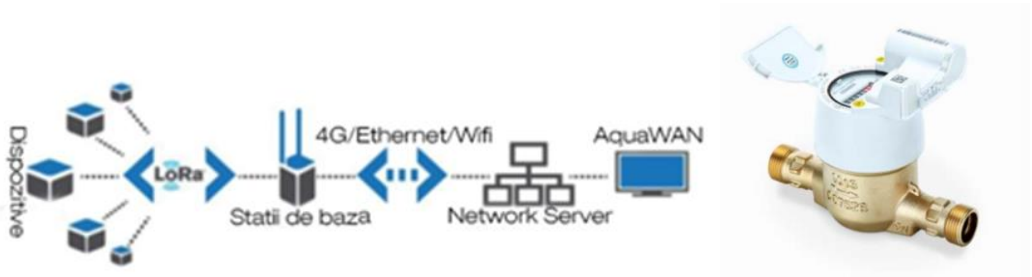

Figura 2.1 Sistemul AquaWAN si contor inteligent de colectare a datelor

- ISTA radio – sistem radio de colectare a datelor senzoriale
  - Colectează date de la senzori de apa, încălzire, gaz, energie electrica
  - Componente:
    - stație radio de colectare,

        - pulsonic 3 radio – permite integrarea dispozitivelor de contorizare cu impuls in sistemul radio
        - optosonic radio – pentru conectarea contoarelor de energie termica
- ISTA M-BUS – sistem de colectare a datelor senzoriale bazat pe o rețea industriala de comunicații tip M-BUS
    - Sistem inteligent si integrat de contorizare pentru: energie termica, energie electrica, apa, gaze
    - Componente:
        - sensonic® II M-Bus – contorizare energie termica
        - domaqua® M-Bus – contorizare apa rece si calda
        - Pulsonic II M-Bus, - contorizare energie electrica si gaze
- AMR radio system – sistem de colectare radio a datelor de consum
    - Colectarea datelor pe cale radio (frecventa 868MHz) prin protocolul Wireless M-BUS (WMBUS)
    - Măsoară consumurile de apa in clădiri: citire cu dispozitiv portabil sau in sistem staționar
    - Componente:
        - Dispozitiv de contorizare: Apator Powogaz; disponibile contoare pentru apa potabila si apa reziduala
        - Radio Smart top module- dispozitiv de transmitere radio a datelor (868MHz, max. 300 m)
        - Smart terminal – dispozitiv de colectare a datelor de tip PDA
        - Modul de comunicație Bluetooth/WMBUS – convertor de protocol
        - Software Inkasent
- Bmeters Hidrodigit, GSD8I
    - Citirea de la distanta a valorilor consumate prin instalarea unui dispozitiv inductiv,
    - Culegerea datelor prin rețea cablata
    - La cerere se poate monta modul care comunica prin protocolul LoRa

Analizând aceste oferte de echipamente si sisteme de monitorizare a consumurilor de apa se pot formula o serie de soluții si caracteristici comune:

- Mijloace de comunicare utilizate:
    - Soluție cablata cu protocol M-BUS/RS485
    - Soluție radio, frecventa 868MHz,
    - Combinație transmisie wireless cu M-Bus
    - Rețea/Protocol LoRa
- Tipuri de senzori de contorizare:
    - Senzori clasici (cu generare de impulsuri) la care se adaugă modul de comunicație
    - Senzori inteligenți cu interfața de comunicație integrata
- Funcții oferite de dispozitivele inteligente de contorizare:
    - Măsurarea debitului si a consumului curent
    - Transmiterea la distanta a informației de consum

    - Colectarea si stocarea locala a datelor pe o anumita perioada
    - Detecția unor tentative de frauda
    - Detecția curgerii inverse
    - Detecția unor scurgeri nejustificate (ex. Datorita unei defecțiuni in instalație)
    - Oprirea de la distanta a debitului
- Sisteme de colectare a datelor:
    - Colectare de la distanta „prin vizitare”
    - Colectare prin rețea cablata
    - Colectare radio instantanee in mai multe puncte
- Integrabilitate:
    - Soluții dedicate, numai pentru contorizarea apei
    - Soluții integrate de măsurare a diverselor utilități „casnice” apa, energie permisa, energie electrica, gaze

B. Dispozitive pentru măsurarea calității apei

Calitatea apei se determina printr-o serie de măsurători fizica-chimice si biologice. De cele mai multe ori determinarea calității apei potabile se realizează printr-un proces manual realizat in laboratoare de specialitate. Din acest motiv monitorizarea este intermitenta si frecventa scăzută a operațiilor de colectare de probe si de analiza (de câteva ori pe zi) nu permite determinarea rapida a unor accidente de contaminare. In cazul unor sisteme automate de monitorizare sunt necesare dispozitive de analiza automate cu capabilitate de transmitere la distanta a datelor.

Exemple de dispozitive de măsurare:

- Aqua TROLL 200 Data Logger – dispozitiv multifuncțional ce permite măsurători de: temperatură, presiune, conductivitate si nivel de apa, cu posibilitatea de stocare a datelor achiziționate si transmiterea acestora către un dispozitiv mobil de colectare si evaluare a datelor.
- Aqua TROLL 400 Multiparameter Probe – dispozitiv multifuncțional pentru măsurarea următorilor parametric de calitate ai apei: Conductivitatea reală și specifică, salinitatea, solidele totale dizolvate, rezistivitatea, concentrație de oxigen dizolvat (RDO® optic), pH, Temperatura, nivel de apa
- Online pH/ORP Sensor (Yanta winmore trade) – un senzor de pH conectabil într-o rețea de tip RS485/Modbus; de la același producător: online turbidity sensor, dissolved oxygen sensor, total suspended solid sensor, conductivity sensor
- Dispozitive on-line oferite de firma Tehno Instrument: analizoare de compuși organici si anorganici, suspensii solide, încărcare bacteriologica (e-coli, alge, etc.), senzori pentru nitrați, nitriți, clor, metale, etc.

C. Senzori si sisteme de măsurare a integrității rețelelor de apa

- Flume Smart Home Water Monitoring System – detectează si transmite mesaj telefonic in cazul apariției unor defecte (scurgeri) in rețeaua de apa; comunicație prin telefonie mobila sau WiFi
- Vivin water sensor – integrat într-un sistem de securizare a unei clădiri; detectează defecțiuni in instalația de apa si transmite mesaj către sistemul de alarmare.

- Ring Alarm Flood and freeze Sensor – transmite mesaj (SMS) si email in caz de detectare a unor scurgeri sau când temperatura tinde către un punct de înghețare a instalației.

D. Senzori clasificați pe baza mărimii fizice măsurate

- Senzori de măsurare a nivelului apei
    - Pentru stațiile automate există următoarele tipuri de sonde de măsurare a nivelului:
    - Incrementale cu plutitor – potrivit măsurătorilor apei cu solide în suspensie, a surselor cu apă murdară, a apelor cu curgere lentă,
    - Cu ultrasunete – potrivit măsurătorilor pentru apa cu impurități solide,
    - De tip radar cu microunde – potrivit atât pe cursurile de apă, cât și in cadrul apelor staționare cum sunt lacurile,
    - Cu senzor de presiune – cel mai potrivit în cazul volumelor mari de apă.
- Senzori pentru măsurarea temperaturii apei
    - Înglobat în cadrul senzorului de nivel,
    - Senzor de temperatură de sine stătător.
- Senzori pentru măsurarea temperaturii aerului exterior
    - Cu rezistenta Pt100 si panouri radiante [9].
- Senzori pentru măsurarea nivelului de precipitații lichide sau solide
    - De tip „tipping bucket” ,
    - Cu camera de acumulare și traductor capacitiv (potrivit instalării pe stații de monitorizare plasate pe luciul apei)
- Sonde pentru măsurarea calității apei
    - Multiparametrice [10] - permit gruparea în diferite configurații a diferitelor categorii de senzori,
    - Senzori specializați pentru o singură mărime fizică măsurată (de ex. de pH, turbiditate, sodiu, clor, nitriți, alge, clorofilă, țiței brut, etc.)

2.2.2. Dispozitive de control

In aceasta categorie intra acele dispozitive de automatizare care permit controlul direct al fluxului de apa, adică robinete, valve si pompe de apa. In cazul robinetelor si valvelor, controlul fluxului de apa se poate face continuu (prin controlul debitului) sau intermitent prin pornirea si oprirea debitului. Ca si elemente de acționare se pot utiliza motoare electrice (pas-cu-pas sau de curent continuu) sau motoare hidraulice sau pneumatice. Pompele sunt acționate mai ales pe cale electrica, utilizând-se in acest scop motoare de curent alternativ (pentru debite mari) sau de curent continuu (pentru acurateţe mai mare).

Exista o foarte mare varietate de dispozitive de acționare pentru controlul debitului de apa. Vom da doar câteva exemple sugestive [11]:

- Smart Water Valve,Wifi Control Water/Gas Shut Off Valve – valve cu acționare prin rețea WiFi si prin voce
- Dome Home Automation Water Shut Off – valve pentru oprirea automata a instalației in caz de scurgeri accidentale; acționare WiFi
- Robinete electrice furnizate de firma Recondi:
- robinete sferice cu acționare prin motoare de curent alternativ (230V, 24Vca) si de curent continuu (24Vcc, 12Vcc)

- robinet ventil cu acționare electrica
- electroventile normal închis sau normal deschis; acționare prin solenoid, 2 stări posibile
- Ventil cu bila 2 cai motorizat ESBE MBA, acționat electric
- Robinete de reglare cu acționare electrica (producător ESC) – acționare On/OFF
- Robinet sferic cu acționare electrica Lyva 2 PN10 (producător Vannes Europ)

3. Sisteme generice pentru stocarea si procesarea datelor de la senzori

3.1 Platforme de tip cloud IoT

Platformele generice de tip cloud IoT pot fi descrise prin arhitectura [12] din Figura 3.1 și includ patru servicii principale: servicii pentru colectarea, stocarea, procesarea datelor și servicii de execuție de operații.

Datele pot proveni din diverse surse, cum ar fi senzori, diferite dispozitive inteligente sau chiar și utilaje industriale sau mașini inteligente. Datele colectate sunt transmise de la dispozitiv către o unitate de tip gateway, utilizând diverse protocoale de rețea, cum ar fi FieldBus, ControlBus, OPC, Message Queue Telemetry Transport (MQTT), Advanced MQTT (AMQTT), HTTP, HTTPS, Constrained Application Protocol (CoAP). Protocoalele de rețea transferă date către un IoT-Hub, care acționează ca o poartă de acces între dispozitiv și cloud. De aici datele pot fi transmise în mai multe direcții.

Datele pot fi direcționate către o bază de date, care ulterior permite analiza avansata (Advanced Analytics) de către o anumită aplicație bazată pe inteligență artificială. Aceasta se numește calea rece (cold path) a procesării datelor.

Serviciul IoT-Hub oferă, de asemenea, capacitatea de procesare rapidă, instantanee, numită Stream Analytics. Aceasta corespunde căii calde (hot path) a analizei și poate oferi rezultate în timp real, dacă este necesar.

Pe baza acestor analize, se pot crea diferite rezultate utile pentru un operator uman, cum ar fi crearea de alarme, detecția anomaliilor, generarea de rapoarte, agregarea și vizualizarea datelor, etc. Deci, generează informații utile pentru utilizator din datele colectate și transmise de dispozitive (edge devices).

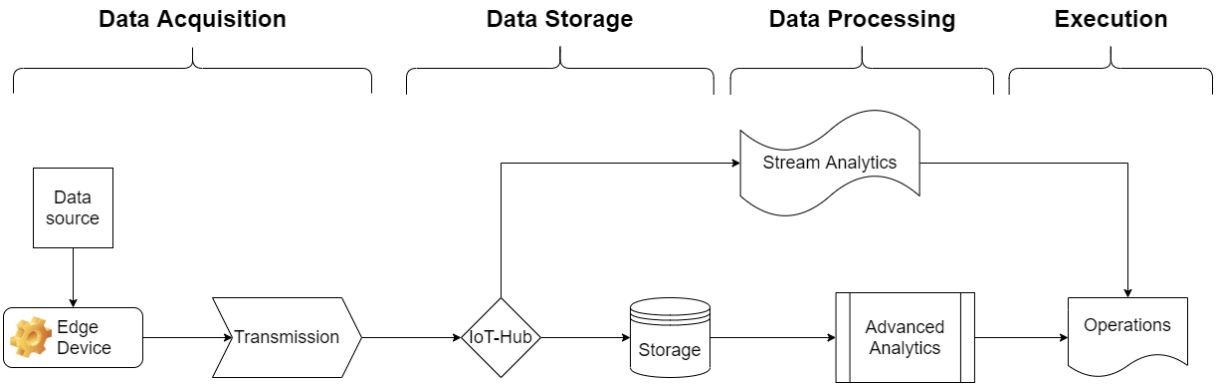

Figura 3.1. Arhitectura generală servicii cloud IoT [12]

În cele ce urmează, va fi descrisa modul in care arhitectura din figura 3.1 se mapeaza peste serviciile cloud IoT ale Microsoft Azure [13], Amazon Web Services [14] și Google Cloud [15].

3.1.1 Microsoft Azure

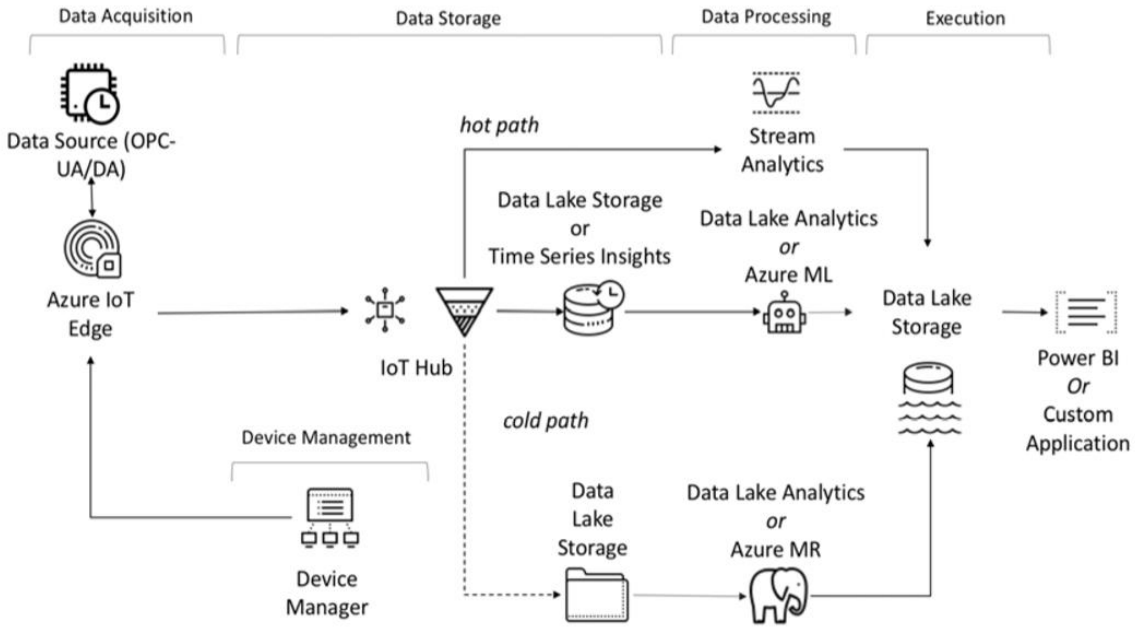

Figura 3.2. Azure IoT

Datele sunt colectate de la senzori și alte dispozitive industriale prin intermediul Azure IoT Edge și prin diferite protocoale de transmisie, cum ar fi MQTT, AMQP, HTTP, OPC, etc., sunt redirecționate către Azure IoT Hub.

IoT-Hub este un strat intermediar pentru înregistrarea și gestionarea dispozitivelor. De aici, datele sunt procesate de Stream Analytics, care este calea fierbinte pentru analiză.

Azure Stream Analytics este un motor de analiză și procesare complexă în timp real de evenimente, conceput pentru a analiza și procesa simultan cantități mari de date, care pot proveni din mai multe surse. Rezultatele analizei sunt stocate în Data Lake Storage.

Datele pot fi stocate în două tipuri de baze de date, una este Time Series Insight și cealaltă este Azure Data Lake Storage. Time Series Insight este capabil să stocheze și să vizualizeze cantități mari de date de tip serii de timp. Azure Data Lake Storage este un depozit foarte scalabil care permite stocarea datelor pentru analiză și afișare ulterioară. Data Lake Analytics este un serviciu de date mari pentru prelucrarea unor cantități imense de date.

ML Analytics este un mediu dedicat analizelor bazate pe învățarea automată avansată, pentru modelare predictivă.

3.1.2 Amazon Web Services (AWS)

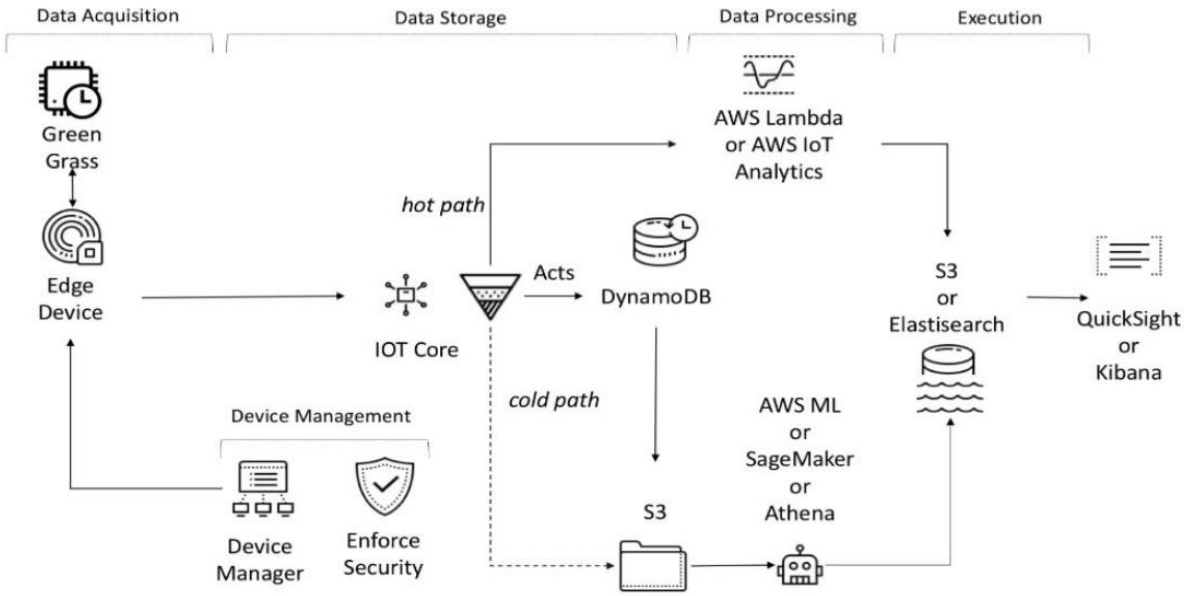

Figura 3.3. AWS IoT

GreenGrass este un software care extinde capacitățile cloud-ului la dispozitivul local și permite dispozitivului să colecteze și să analizeze date mai aproape de sursa de informații (edge computing).

Există mai multe opțiuni pentru stocarea datelor și aici, în funcție de tipul de date pe care dorim să le stocăm, putem utiliza baza de date S3 pentru a stoca seriilor de timp sau DynamoDB pentru stocarea datelor în general. Pentru procesarea rapidă a datelor, putem utiliza platforma serverless AWS Lambda sau AWS IoT Analytics. Pentru analize mai avansate, AWS oferă mai multe opțiuni pentru utilizator: AWS ML, SageMaker, Athena. Rezultatele analizelor avansate sunt fie stocate în baze de date de tip S3 sau Elastic Search sau pot fi legate direct de serviciile QuickSight sau Kibana, care sunt responsabile pentru procesare și vizualizare rapidă.

3.1.3 Google Cloud IoT

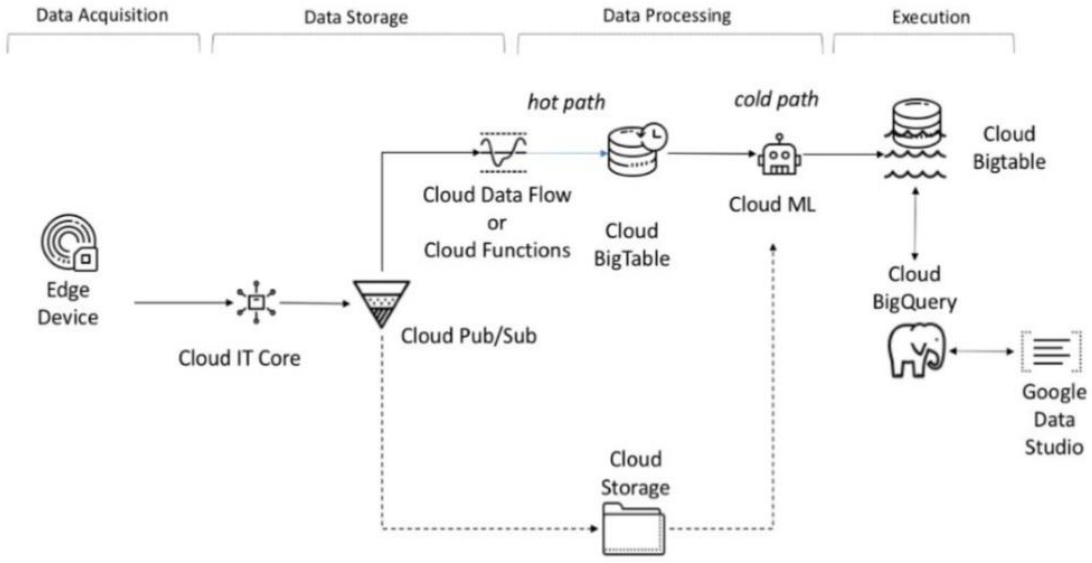

Figura 4.4. Google IoT

Google IoT Core acceptă numai protocoalele MQTT și HTTPS. În acest caz, componenta numită Stream Analytics este compusă din Cloud Data Flow sau Cloud Functions. Cloud Bigtable poate fi utilizat pentru a stoca date de tip serii de timp sau de tip eveniment, iar Cloud Storage poate fi utilizat pentru a stoca date de tip obiect. Cloud BigQuery și Cloud ML pot fi utilizate pentru analiza batch a datelor. Google Data Studio este responsabil pentru vizualizarea datelor.

3.2 Servicii standard pentru managementul datelor provenite de la senzori

3.2.1 OGC Sensor Observation Service (SOS)

Standardul SOS este un standard internațional pentru administrarea datelor provenite de la senzori. El a fost dezvoltat de OGC (Open Geospatial Consortium) [16]. SOS își propune să standardizeze formatul datelor obținute prin măsurare, provenite de la diferite surse (senzori) astfel încât să existe o interfață comună, reducând numărul de implementări necesare pentru fiecare format în parte, la un singur format standard. Datele ajung astfel în formatul prestabilit la serverul SOS, după care acestea pot fi preluate de către clienți folosind aceeași implementare standard.

Operațiile care pot fi realizate în SOS sunt următoarele:

- ***GetCapabilities***, pentru a cere auto-descrierea serviciului.
- ***GetObservation***, pentru a cere date generale despre senzori.
- ***DescribeSensor*****,** pentru a cere date specifice despre un senzor, codificate cu SensorML.
- ***GetFeatureOfInterest***, pentru a cere informații despre obiectivul măsurătorilor.
- ***GetObservaitonById***
- ***InsertSensor***, pentru publicarea de noi senzori.
- ***UpdateSensorDescription***, pentru actualizarea descrierii senzorilor.
- ***DeleteSensor***, pentru ștergerea unui senzor.
- ***InsertObservation***, pentru inserarea de observații.

- ***InsertResultTemplate***, pentru inserarea de template-uri care vor fi folosite la inserarea de rezultate.
- ***InsertResult***, pentru a încărca rezultate neprocesate conform structurii precizate în template.
- ***GetResultTemplate***, pentru a obține structura rezultatului.
- ***GetResult***, pentru a descărca rezultate neprocesate conform structurii precizate în template.

Clientul SOS pune la dispoziție diferite servicii web prin care se face accesul la date. Unul dintre acestea este GetCapabilities (Preluare Capacități) care oferă utilizatorului toate informațiile despre senzorii existenți, coordonatele acestora, tipurile de date măsurate, interfețele disponibile, perioadele între care sunt disponibile datele de la senzori sau fenomenele observate.

Operația InsertSensor (Inserare Senzor) este folosită pentru crearea unui nou senzor în baza de date SOS. Datele trimise prin această operație sunt o secvență alfanumerică de identificare a senzorului, un nume scurt și unul lung (opțional), un indicator de ofertă, denumirea procedurii părinte (rețeaua de senzori principală), regiunea de interes, poziția senzorului cu latitudine, longitudine și înălțime, denumirea proprietății observate (ex: temperatura) și meta date, precum specificarea unei măsurări de tip text, valoare de adevăr sau caracteristica geometrică.

Operația InsertResultTemplate (Inserare Șablon Rezultat) dă posibilitatea creării unui șablon folosit mai departe la inserarea de rezultate propriu-zise. Șablonul este format din identificatorul șablonului, identificatorul de ofertă descris la operația de inserare senzor, regiunea de interes, numele șablonului, coordonatele caracteristicii măsurate, toate acestea făcând parte din șablonul observației. A doua componentă este șablonul rezultatului, acesta fiind format din timpul fenomenului, numele proprietății măsurate și unitatea de măsură a proprietății măsurate.

Operația InsertResult (Inserare Rezultat) se folosește pentru a insera valorile măsurate de către senzori în baza de date. Parametrii acestui serviciu web sunt identificatorul șablonului descris mai sus și valoarea rezultatului.

Preluarea de date de la serviciul SOS se face prin operația GetResult (Preluare Rezultat). Serviciul web primește ca parametri oferta dorită și proprietatea observată dorită, iar acesta va returna valorile în formatul descris la operația „inserare rezultat”.

Clientul SOS poate realiza persistența bazei de date folosind modelul de date specificat in standard.

Exista mai multe implementari open source SOS, inclusiv una speciala pentru date hidrografice (de la 52North).

3.2.2 OGC SensorThings API

API-ul OGC SensorThings [17] oferă o modalitate deschisă și unificată de a interconecta dispozitivele, datele și aplicațiile Internet of Things (IoT) de pe web. API-ul OGC SensorThings oferă două funcționalități principale. Funcționalitatea Sensing oferă o modalitate standard de gestionare și colectare a observațiilor și metadatelor din sistemele de senzori IoT eterogene. Funcționalitatea Tasking oferă o modalitate standard de parametrizare - numită și tasking - a dispozitivelor IoT cum ar fi senzori și actuatori individuali, platforme in-situ compuse (de tip consumator / comercial / industrial / orașe inteligente), dispozitive mobile și purtabile sau drone, sateliți, vehicule conectate și autonome etc. API-ul urmează principiile REST, codificarea JSON, protocolul OASIS OData și convențiile URL. De asemenea, are o extensie MQTT care permite utilizatorilor / dispozitivelor să publice și să trimită actualizări de pe dispozitive și poate utiliza CoAP în plus față de HTTP.

Un dispozitiv sau sistem IoT este modelat ca un Thing. Un Thing are un număr arbitrar de Locations (inclusiv 0 Lacationa) și un număr arbitrar de Datastreams (inclusiv 0 Datastreams). Fiecare Datastream observă o ObservedProperty cu un Sensor și are mai multe Observations colectate de Sensor. Fiecare Observation observă un anumit FeatureOfInterest.

Exista mai multe implementari open source ale acestui standard:

- Whiskers [18]
- GHOST[19]
- Mozilla STA [20]
- 52North STA [21]

4. Procesarea datelor si detecția anomaliilor

4.1 Formatul datelor

Alegerea formatului optim pentru stocarea datelor poate îmbunătăți performanța modelelor și procesarea datelor.

Formatele cele mai populare in domeniul inteligenței artificiale sunt:

- Comma-separated values
- XLSX
- ZIP
- Plain Text (txt)
- JSON
- XML
- HTML
- Images
- Hierarchical Data Format
- PDF
- MP3
- MP4

Pentru a avea un sistem omogen, în care putem trata uniform orice fel de format al datelor de intrare, formatul comun, folosit intern se numește Numpy NDArray [22]. Această structură de date reprezintă o matrice multidimensională, omogenă de elemente de dimensiuni fixe. Am ales această reprezentare a datelor, pentru că toate formatele specificate mai sus pot fi stocate în această structură de date și e acceptată de către toate librăriile populare de inteligență artificială.

4.2 Metode de preprocesare a datelor

Există trei categorii mari de metode de preprocesare:

- Curățarea datelor
- Transformarea datelor
- Reducerea datelor

4.2.1. Curățarea datelor

Scopul curățării datelor este de a oferi seturi de date simple, complete și clare pentru învățarea automată.

Date lipsă

De multe ori trebuie să lucrăm cu date de intrare în care lipsesc anumite date, ceea ce poate strica performanța modelului de învățare automată. Pentru a rezolva această problemă avem mai multe posibilități:

- Ștergerea rândurilor cu valori lipsă
- Completarea valorile lipsă folosind media, mediana sau valori implicite
- Utilizarea algoritmilor care acceptă valorile lipsă
- Predicția valorilor lipsă

Date zgomotoase

Când datele sunt colectate, oamenii tind să comită greșeli, iar instrumentele tind să fie inexacte, astfel încât datele colectate au o anumită eroare. Această eroare este denumită zgomot.

Zgomotul creează probleme pentru algoritmii de învățare automată, deoarece dacă nu sunt antrenați corect, algoritmii pot considera că zgomotul este un model și pot începe să se generalizeze din acesta, ceea ce, desigur, este nedorit.

Pentru reducerea zgomotului avem mai multe posibilități:

- Reducerea dimensionalității folosind Analiza Componentelor Principale (PCA – Principal Component Analysis) [24] sau Analiza Discriminantă Liniară (LDA – Linear Discriminant Analysis) [61]
- Tehnici de curățare și normalizare specifice domeniului (de exemplu folosind filtru trece bandă în cazul seriilor de timp, eliminarea termenilor de tip stop word, algoritmi de reducerea zgomotului pentru imagini, etc.)
- Folosirea sistemului bazat pe voting și averaging pentru a ține cont de zgomotul sistematic și pentru a îmbunătăți generalizarea
- Combinarea mai multor modele cu Boosting, în timp ce un singur model poate fi susceptibil la zgomot, un ansamblu minimizează impactul acestuia

4.2.2. Transformarea datelor

Normalizare/Scalare

Scopul normalizării este de a schimba valorile coloanelor numerice din setul de date la o scară comună, fără a distorsiona diferențele în intervalele de valori. Aici putem vorbi de metode care setează media datelor la 0 și deviația standard la 1, metode care scalează intervalul de date la [0,1], metode care micșorează/întinde un vector (un rând de date poate fi văzut ca un vector D-dimensional) la o sferă unitară, etc.

Selecția caracteristicilor

Atunci când construim un model de învățare automată în viața reală, este foarte rar ca toate variabilele din setul de date să fie utile pentru a crea modelul. Adăugarea de variabile redundante reduce capacitatea de generalizare a modelului și poate reduce, de asemenea, precizia generală a unui clasificator. În plus, adăugarea din ce în ce mai multe variabile la un model crește complexitatea generală a modelului.

Scopul selecției caracteristicilor în învățarea automată este de a găsi cel mai bun set de caracteristici care să permită construirea unor modele utile ale fenomenelor studiate.

Există o mulțime de moduri în care putem selecta caracteristicilor importante, dar majoritatea metodelor de selectare a caracteristicilor pot fi împărțite în trei categorii majore:

- Metode bazate pe filtre: prin specificarea unor praguri sau metrici. Un exemplu de astfel de metrică ar putea fi corelația, varianta, diferența absolută medie, raport de dispersie, scorul lui Fisher, testul Chi-pătrat, etc.
- Metode bazata pe căutare: metodele bazate pe căutare consideră selectarea unui set de caracteristici ca o problemă de căutare, de exemplu utilizând metoda numita Eliminarea Recursiva a Caracteristicii [23]
- Metode încorporate: metodele încorporate utilizează algoritmi care au încorporate metode de selectare a caracteristicilor. De exemplu, Lasso și Random Forest au propriile metode de selectare a caracteristicilor.

Discretizare

Variabilele numerice de intrare pot avea o distribuție foarte înclinată sau nestandardizată. Acest lucru ar putea fi cauzat de valori anormale în date, distribuții multimodale, distribuții exponențiale și multe altele.

Mulți algoritmi de învățare automată au performanțe mai bune atunci când variabilele de intrare numerică au o distribuție standard. Discretizarea oferă o modalitate automată de a schimba o variabilă de intrare numerică pentru a avea o distribuție diferită.

4.2.3. Reducerea datelor

Reducerea dimensionalității

Dacă avem un număr mare de dimensiuni în spațiul caracteristic al datelor de intrare poate însemna că volumul spațiului respectiv este foarte mare și, la rândul său, punctele pe care le avem în acel spațiu (rânduri de date) reprezintă adesea un eșantion mic și nereprezentativ. Acest lucru poate avea un impact dramatic asupra performanței algoritmilor de învățare automată, denumit în general „blestemul dimensionalității”. Prin urmare, este adesea de dorit să reducem numărul de caracteristici de intrare.

Metode pentru reducerea dimensionalității:

- Metodele de filtrare - folosind corelația dintre caracteristică și variabila țintă, pentru a selecta unui subset de caracteristici de intrare care sunt cele mai predictive. Exemple includ corelația lui Pearson și testul Chi.
- Factorizarea matricei - tehnicile din algebra liniară pot fi utilizate pentru reducerea dimensionalității. În mod specific, metodele de factorizare a matricei pot fi utilizate pentru a reduce o matrice de date în părțile sale constitutive. Exemplele includ descompunerea valorilor proprii și descompunerea valorii singulare. Cea mai comună metodă de reducerea dimensionalității este analiza componentelor principale sau PCA [24] pe scurt.
- Învățare multiplă - aceste tehnici sunt folosite pentru a crea o proiecție de dimensiuni reduse, de multe ori în scopul vizualizării datelor. Proiecția este concepută atât pentru a crea o reprezentare dimensională redusă a setului de date, păstrând în același timp structura sau relațiile evidente din date. Exemple de algoritmi: Multidimensional Scaling[25], t-distributed Stochastic Neighbor Embedding (t-SNE)[26], Sammons Mapping [27], Kohonen Self-Organizing Map (SOM) [28], etc.

Reducerea numerozității

Reducerea numerozității este o tehnică de reducere a datelor care înlocuiește datele originale printr-o formă mai mică de reprezentare a datelor. Există două tehnici pentru reducerea numerozității - metode parametrice și non-parametrice.

Pentru metodele parametrice, datele sunt reprezentate folosind un model. Modelul este utilizat pentru a estima datele, astfel încât numai parametrii să fie stocați, în loc de date reale. Pentru crearea unor astfel de modele sunt utilizate metode de regresie și Log-Linear.

Metode non-parametrice sunt utilizate pentru stocarea reprezentărilor reduse ale datelor, incluzând histograme, grupare, eșantionare și agregare de date.

4.3. Detecția anomaliilor în sisteme cyber-fizice

„Anomalie" este un termen general care poate acoperi o multitudine de cauze care scot un sistem din condițiile sale normale de funcționare [63][64]. Datorită anomaliilor, seturile de date de monitorizare vor include valori aberante. Termenul „outlier" a fost folosit inițial în domeniul statisticilor și este definit în [65] ca o observație care este incompatibilă cu setul de date de care aparține. Chiar dacă valorile anormale din seturile de date sunt efectul anomaliilor în funcționarea sistemului, în multe cazuri, termenii „detecție de outlier" și „detecție de anomalii" sunt utilizați cu același sens.

Există un consens general asupra faptului că detecția anomaliilor este un subiect vast, care are aplicații în diverse domenii și că nu există o soluție generală la problema de detecție a anomaliilor. Soluțiile variază în funcție de caracteristicile și particularitățile fiecărei probleme. În acest context, a existat un mare interes pentru clasificarea anomaliilor și a metodelor de detecție.

Majoritatea clasificărilor tehnicilor de detectare a anomaliilor prezentate într-o serie de articole[66][67][68][69] oferă o viziune unilaterală a tehnicilor existente, o perspectivă bazată pe un singur criteriu.

4.3.1. Criterii de clasificare a metodelor de detecție a anomaliilor

Pe baza literaturii existente, au fost identificate următoarele criterii care pot fi folosite pentru clasificarea metodelor de detecție a anomaliilor:

Natura anomaliei

Folosind acest criteriu putem identifica următoarele clase:

- anomalie într-un punct - un punct (din colecția atributelor unui sistem la un moment dat) care este foarte diferit de punctele considerate normale; tehnicile statistice și de distribuție normală sunt adecvate pentru acest tip de detecție a anomaliilor; în acest caz avem două subclase:
    - unele tehnici iau în considerare relația dintre un punct și toate celelalte puncte din set
    - altele consideră relația (de exemplu, distanța) dintre un singur punct și membrii clasei de puncte detectate anterior; aici sunt combinate tehnicile de grupare împreună cu metricile bazate pe distanță pentru detecția anomaliilor
- anomalie contextuală - o anomalie mai complexă care ia în considerare relația dintre punctul analizat și vecinii săi; tehnicile utilizate aici iau în considerare schimbările din setul de date; o sarcină specială aici este de a determina în mod automat sau într-un mod iterativ lungimea optimă a ferestrei care definește contextul unui punct
- anomalie într-o secvență de date - nu consideră valoarea unui punct individual, ci forma multidimensională a secvențelor de puncte ordonate în timp; exemplul tipic aici este recunoașterea secvențelor anormale în semnalele ECG; tehnicile utilizate aici pot fi construite pe trăsături extrase din analiza semnalului (de exemplu FFT, corelație etc.) sau metode care pot detecta asemănările și disparitățile în anumite secvențe

Natura datelor analizate

Acest criteriu conține o serie de subcriterii:

- Numărul de atribute care caracterizează un punct din setul de date:
  - Puncte univariate - caracterizat printr-o singură valoare a parametrului
  - Puncte multivariabile - un punct caracterizat de un număr de atribute sau valori ale parametrilor; atributele pot fi numerice, textuale sau mixte.
- Corelația dintre puncte:
  - corelație statistică - puncte de date care respectă anumite reguli statistice sau probabilistice;
  - corelație în timp - datele reprezintă serii de timp în care un punct de date este corelat cu eșantioanele anterioare ale aceluiași parametru; filtrarea, medierea în mișcare sau tehnicile de auto-regresie pot fi aplicate pentru detectarea anomaliilor; există, de asemenea, tehnici care încearcă să identifice modelul sistemului care generează date anterioare și să compare valoarea estimată (calculată cu modelul) cu valoarea reală; o anomalie este detectată atunci când diferența depășește un prag predefinit;
  - corelația spațială - este similară cazului precedent cu diferența că această corelația este între punctele învecinate spațial; acest caz este tipic pentru datele colectate de la rețelele de senzori în care există o conexiune fizică între senzorii învecinați;
  - corelație funcțională - în acest caz există o relație funcțională între diferiții parametri măsurați într-un sistem (de exemplu, curent, tensiune și consum de energie într-o rețea de distribuție a energiei electrice); corelația poate fi derivată din legile fizice care guvernează procesul sau poate fi dedusă prin calcule inter-corelaționale;

Natura principiilor de detectare

Metodele de detecție a anomaliilor pot fi bazate pe tehnici generice împrumutate din diferite domenii de cercetare / știință: statistică, inteligență artificială, exploatarea datelor, recunoașterea modelului, teoria sistemelor, prelucrarea semnalelor și multe altele. Mai jos sunt enumerate tehnici adoptate cu succes pentru detecția anomaliilor:

- Metode statistice - considerăm că datele colectate ar trebui să aibă o distribuție de probabilități dată; o valoare mai mare (date anormale) depășește un interval considerat normal (de exemplu, valoarea medie +/- 2 * deviație standard); metode mai complexe îmbunătățesc precizia prin examinarea variațiilor sezoniere din setul de date; aceste metode sunt tipice pentru colectarea datelor financiare;
- Metode de inteligență artificială - includ diferite tehnici de căutare, clasificare, grupare și alte tehnici, grupate sub denumirea generică de inteligență artificială; tipic pentru această categorie de metode de detectare este că normele de discriminare între valori normale și anormale nu sunt date; algoritmul încearcă să identifice anumite reguli (discriminatorii) pentru recunoașterea anomaliilor; în funcție de metoda de învățare există două abordări:
  - Metoda de învățare:
    - tehnici supravegheate - regulile sau discriminatorii sunt învățate din exemple pozitive și negative (eșantioanele de date sunt etichetate ca fiind normale sau anormale);
    - tehnici semi-supervizate - numai comportamentul normal poate fi învățat din exemple pozitive;

        - tehnicile nesupravegheate - procesul de învățare nu poate lua în considerare unele anomalii specificate a priori.
    - Metode aplicate:
        - gruparea și clasificarea - metoda de detectare consideră că un outlier nu este suficient de "apropiat" clusterelor de date definite sau determinate; există o varietate de metode care iau în considerare metrici diferite pentru a determina apropierea datelor de un cluster dat (de exemplu, măsurarea distanței sau densității locale)
        - forma și recunoașterea modelului - aici nu datele individuale, ci forma semnalului (secvența eșantioanelor) face diferența între o stare normală sau o stare anormală; SVM (Support Vector Machine), rețele neuronale sau lanțuri Markov sunt instruite pentru a face diferența între forma normală și cea anormală
        - potrivirea șirului - în acest caz evoluția unui semnal este recodificată cu un număr finit de simboluri (de exemplu, caractere), iar problema identificării unei anomalii este redusă la problema detectării unui șir care nu se potrivește cu șirurile considerate normale
- Teoria sistemelor și metodele de procesare a semnalelor - aceste metode [6][9] lucrează cu serii de timp sau cu date spațial-temporale; unele dintre metode utilizează tehnici de filtrare mai mult sau mai puțin simple pentru a detecta valori excepționale; alte metode încearcă să modeleze comportamentul sistemului generând datele și compară valoarea estimată (bazată pe model) cu datele furnizate efectiv:
    - min-max, prag - acestea sunt cele mai simple forme de detectare a anomaliilor, unde valorile minime, maxime sau prag sunt determinate dintr-o secvență considerată normală; valorile care depășesc aceste limite sunt considerate anomalii;
    - filtrarea - prin compararea unui semnal cu versiunea filtrată folosind un filtru trece-jos, poate indica o anomalie; ideea din spatele metodei este aceea că o valoare mai mare (de exemplu un zgomot în semnal) nu păstrează "continuitatea" semnalului;
    - modelare - aici, folosind tehnici de identificare a sistemului, este construit un model al procesului pentru generare a datelor; acest model este folosit pentru a prezice următoarele valori; în cazul în care există o eroare semnificativă între datele colectate și cea anticipată, concluzia este că avem o anomalie; există diferite tehnici utilizate în acest scop, cum ar fi: auto-regresia, auto-corelația, corelarea multivariată etc.

Natura domeniului aplicației

Uneori, tipul de metodă de detecție a anomaliilor utilizat într-un caz dat este determinat (cel puțin parțial) de domeniul aplicației. De exemplu, în cazul datelor financiare, se recomandă utilizarea metodelor statistice sau, în cazul aplicațiilor industriale, se utilizează metode înrudite cu serii de timp. Prin urmare, un criteriu de clasificare pentru metodele de detecție a anomaliilor este domeniul de aplicare. Iată câteva dintre domeniile în care detecția anomaliilor joacă un rol important: date economice și financiare; sisteme cyber-fizice și procese industriale, date privind mediul, prognoza meteo; date geo-spațiale, diagnostice medicale, calculatoare și rețele de calculatoare (detectarea virușilor, detectarea unui cod rău intenționat).

În funcție de domeniul aplicației, criteriile de calitate utilizate pentru selectarea celei mai bune metode de detecție pot fi diferite; în unele domenii, câteva fals pozitive nu sunt critice (de exemplu, prognoza meteo, date privind mediul), dar în altele, precizia este esențială (de exemplu domeniul medical, procesele industriale). În unele domenii detectarea trebuie efectuată on-line (de exemplu, sisteme

cibernetice-fizice, procese industriale), restricționând fereastra de căutare și timpul de execuție disponibil; în alte cazuri, procesarea off-line se face fără restricții de timp (de exemplu, date financiare, diagnostice medicale), permițând o implementare mai precisă a metodei de detectare însă cu timp de execuție mai îndelungată.

Observații finale privind criteriile de clasificare metodelor de detecție a anomaliilor

Multitudinea metodelor de detectare a anomaliilor prezente în literatura științifică nu poate fi clasificată printr-un singur criteriu sau dintr-un singur punct de vedere. Trebuie aplicate mai multe criterii pentru a poziționa o anumită metodă pe "harta" detecției anomaliilor. Dimensiunile acestei hărți sunt după cum urmează:

- tipul de anomalie abordat printr-o metodă dată (de exemplu, singular, contextual, legate de formă etc.)
- felul datelor analizate (de exemplu, date statistice, serii de timp, date spațio-temporale etc.)
- contextul teoretic al metodei (de exemplu, teoria sistemului, statisticile, inteligența artificială, teoria corzilor etc.)
- domeniul de aplicare (de exemplu, economie, industrial, mediu, medical, etc.)

În fiecare dimensiune majoră există subdiviziuni, așa cum au fost specificate mai sus. Având în vedere această hartă, este mult mai ușor pentru un dezvoltator să abordeze o problemă analitică practică, să selecteze cele mai bune metode și să facă comparații de eficiență.

4.3.2. Clasificarea metodelor de detecție a anomaliilor după natura metodei folosite

Tehnici bazate pe statistici

Aceștia sunt cei mai vechi algoritmi pentru detectarea anomaliilor și unii dintre aceștia sunt aplicabili numai pentru seturi de date uni/bidimensionale sau cu dimensiuni reduse. În acest caz presupunerea este că instanțele de date normale apar în regiuni cu probabilitate ridicată ale unui model stohastic, în timp ce anomaliile apar în regiunile cu probabilitate scăzută. Exemple de metode de detecție a anomaliilor pentru această categorie ar fi metoda lui Grubbs [30](numit Extreme Studentized Deviate) care se bazează pe valoare medie și deviație standard, metode bazate pe histogramă sau tehnici bazate pe modelul Gaussian, presupunând că datele sunt generate dintr-o distribuție gaussiană.

Avantajul tehnicilor statistice este că, dacă ipotezele care iau în considerare distribuția datelor sunt adevărate, tehnicile statistice pot oferi o soluție statistic justificabilă pentru detectarea anomaliilor. Dezavantajul este că aceste tehnici de detecție a anomaliilor au la bază presupunerea că datele sunt generate dintr-o anumită distribuție, iar dacă această presupunere este falsă, rezultatele vor fi greșite.

Detecție bazată pe vecini

Ideea de bază a metodelor de detecție a anomaliilor bazate pe vecin este de a identifica valorile aberante comparându-le cu informațiile din vecinătate. Cel mai reprezentativ algoritm pentru această categorie este k-Nearest-Neighbours (kNN) [31], caz în care scorul de anomalie este definit ca distanța medie (sau distanța ponderată) până la vecinii săi. Alți algortimi din această categorie sunt Local Outlier Factor (LOF) [32], Connectivity-based Outlier Factor (COF) [33], Local Correlation Integral (LOCI) [34], Approximate Local Correlation Integral (aLOCI), etc.

Avantajele tehnicilor de detecție a anomaliilor bazate pe vecini sunt că aceste metode sunt nesupravegheate prin natura lor. Acestea sunt bazate exclusiv pe date de intrare, nu au nevoie de etichete, pot fi utilizate pe diferite tipuri de date, trebuie doar să se definească o măsură de distanță adecvată. Există și unele dezavantaje, cum ar fi dacă datele au instanțe normale care nu au suficienți vecini

apropiați, tehnica eșuează. De asemenea, complexitatea de calcul a fazei de testare este o provocare semnificativă, deoarece implică calcularea distanței pentru fiecare instanță de testare cu toate instanțele aparținând fie datelor de testare, fie datelor de instruire.

Detecție bazată pe sub-spații

În acest caz, presupunerea este că datele pot fi încorporate într-un sub-spațiu cu dimensiuni inferioare, în care instanțele și anomaliile normale par semnificativ diferite.

Mai multe tehnici folosesc analiza componentelor principale (PCA – Principal Component Analysis) [24] pentru proiectarea datelor într-un spațiu cu dimensiuni mai mici. O astfel de tehnică analizează proiecția fiecărei instanțe de date de-a lungul componentelor principale cu varianță mică. O instanță normală care satisface structura de corelație a datelor va avea o valoare scăzută pentru astfel de proiecții, în timp ce o instanță anormală care se abate de la structura de corelație va avea o valoare mare.

Avantajele tehnicilor de detecție a anomaliilor bazate pe sub-spații sunt următoarele:

- Tehnicile bazate pe sub-spații efectuează automat reducerea dimensionalității și, prin urmare, sunt potrivite pentru manipularea seturilor de date cu dimensiuni ridicate. Mai mult, ele pot fi folosite și ca etapă de preprocesare urmată de aplicarea oricărei tehnici de detectare a anomaliilor existente în spațiul transformat.
- Tehnicile bazate pe subspațiu pot fi utilizate într-un cadru nesupravegheat.

Dezavantajele tehnicilor de detecție a anomaliilor bazate pe sub-spații sunt următoarele:

- Tehnicile bazate pe sub-spații sunt utile numai dacă instanțele normale și anormale sunt separabile în încorporarea dimensională inferioară a datelor.
- Tehnicile bazate pe sub-spații au de obicei o complexitate de calcul ridicată.

Detecție bazată pe clusterizare

Ideea principală din spatele utilizării clusterizării pentru detectarea anomaliilor este de a învăța reprezentarea datelor normale în datele deja disponibile (date de antrenare) și apoi utilizarea acestor informații pentru a indica dacă un punct este anormal sau nu atunci când sunt furnizate date noi (test).

Avantajele tehnicilor de detecție a anomaliilor bazate pe cluster sunt că pot funcționa într-un mod nesupravegheat și faza de testare în aceste tehnici e rapidă, deoarece fiecare instanță de date trebuie comparată cu un număr mic de clustere.

Cele mai importante dezavantaje sunt că sunt foarte complexe din punct de vedere al calculului și sunt eficiente numai dacă anomaliile nu se formează clustere semnificative între ele.

Detecție bazată pe clasificare

Un clasificator care poate distinge între clasele normale și anormale poate fi învățat în spațiul datelor de intrare.

Pe baza etichetelor disponibile pentru faza de instruire, tehnicile de detectare a anomaliilor bazate pe clasificare pot fi grupate în două mari categorii: tehnici de detectare a anomaliilor multi-clase și o clasă.

Tehnicile de detectare a anomaliilor bazate pe clasificare multi-clase presupun că datele de instruire conțin instanțe etichetate aparținând mai multor clase normale. Astfel de tehnici de detecție a anomaliilor învață un clasificator pentru a distinge între fiecare clasă normală de restul claselor. O instanță de testare este considerată anormală dacă nu este clasificată normal de niciunul dintre clasificatori. Unele tehnici din această subcategorie asociază un scor de încredere cu predicția făcută de clasificator. Dacă niciunul dintre clasificatori nu este sigur în clasificarea instanței de testare ca normal, instanța este declarată anomală.

Tehnicile de detectare a anomaliilor bazate pe o clasificare cu o clasă presupun că toate instanțele de instruire au o singură etichetă de clasă. Astfel de tehnici învață o graniță discriminativă în jurul instanțelor normale folosind un algoritm de clasificare de o clasă, de exemplu, SVM-uri de o clasă, Kerneli Discriminanți Fisher de o clasă etc.

Avantajele tehnicilor bazate pe clasificare sunt următoarele:

- Tehnicile bazate pe clasificare, în special tehnicile multi-clase, pot folosi algoritmi puternici care pot distinge între instanțele aparținând diferitelor clase.
- Faza de testare a tehnicilor bazate pe clasificare este rapidă, deoarece fiecare instanță de testare trebuie comparată cu modelul pre-calculat.

Dezavantajele tehnicilor bazate pe clasificare sunt următoarele:

- Tehnicile bazate pe clasificare multi-clasă se bazează pe disponibilitatea etichetelor exacte pentru diferite clase normale, ceea ce adesea nu este posibil.
- Tehnicile bazate pe clasificare atribuie o etichetă fiecărei instanțe de test, care poate deveni și un dezavantaj atunci când se dorește un scor semnificativ de anomalie pentru instanțele de test.

Detecție bazată pe rețele neuronale (Deep Anomaly Detection)

Abordările bazate pe rețele neuronale (sau DAD – Deep Anomaly Detection) sunt, în general, non-parametrice și bazate pe modele, generalizează bine la tiparele nevăzute și sunt capabile să învețe limite complexe de clasă.

Există trei abordări principale pentru a rezolva problema detectării anomaliilor utilizând rețele neuronale, metode neuronale supravegheate, semi-supravegheate și metode neuronale nesupravegheate.

Metode neuronale supravegheate

Rețelele neuronale supravegheate utilizează clasificarea datelor pentru a conduce procesul de învățare. Rețelele neuronale folosesc etichetele pentru a regla greutățile și pragurile pentru a se asigura că poate clasifica corect datele de intrare. În acest caz, perceptronii multi-strat sunt utilizați pentru a învăța clasele normale și anormale. Dacă căutăm anomalii în imagini, atunci pot fi utilizate rețele neuronale convoluționale. Pentru a găsi anomalii în diferite texte, sunt disponibile diferite tipuri de rețele neuronale recurente.

Avantajele tehnicilor DAD supravegheate sunt următoarele:

- Metodele DAD supravegheate sunt mai precise decât modelele semi-supravegheate și nesupravegheate
- Faza de testare a tehnicilor bazate pe clasificare este rapidă, deoarece fiecare instanță de testare trebuie comparată cu modelul precalculat.

Dezavantajele tehnicilor DAD supravegheate sunt următoarele:

- Tehnicile supravegheate folosind mai multe clase necesită etichete precise pentru diferite clase normale și instanțe anormale, ceea ce adesea nu este disponibil.
- Tehnicile supravegheate DAD nu reușesc să separe datele normale de datele anormale dacă spațiul caracteristic este extrem de complex și neliniar.

Metode neuronale semi-supravegheate

Etichetele instanțelor normale sunt mult mai ușor de obținut decât anomaliile, ca rezultat, tehnicile DAD semi-supravegheate sunt adoptate pe o scară mai largă. Aceste tehnici folosesc etichetele existente de clasă unică (în mod normal pozitivă) pentru a separa valorile neobișnuite. Un mod obișnuit

de a utiliza aceste rețele în detectarea anomaliilor este de a-i antrena într-un mod semi-supravegheat pe date fără anomalii. Cu suficiente exemple de antrenare, aceste rețele, pentru o clasă normală ar produce eroare scăzuta și eroare ridicata în cazul evenimentelor neobișnuite.

Avantajele tehnicilor DAD semi-supravegheate sunt următoarele:

- Rețelele contradictorii generative (GAN – Generative Adversarial Network) instruite în modul de învățare semi-supravegheat au obținut rezultate foarte bune, chiar și cu foarte puține date etichetate.
- Utilizarea datelor etichetate (de obicei de o clasă) poate produce o îmbunătățire considerabilă a performanței față de tehnicile nesupravegheate.

Metode neuronale nesupravegheate

Rețelele supravegheate necesită un set de date pre-clasificat pentru a permite învățarea. Dacă această pre-clasificare nu este disponibilă, atunci este de dorit o rețea neuronală nesupravegheată. Rețelele neuronale nesupravegheate conțin noduri care concurează pentru a reprezenta porțiuni din setul de date.

După cum sugerează și numele, acest tip de învățare se face fără supraveghere, fără etichete. În timpul instruirii acestor rețele, în cadrul învățării nesupravegheate, vectorii de intrare care sunt similari sunt combinați pentru a forma clustere. Când se aplică un nou tipar de intrare, atunci rețeaua neuronală dă un răspuns de ieșire indicând clasa căreia îi aparține tiparul de intrare. Nu exista feedback din partea mediului cu privire la etichete corecte sau incorecte. Prin urmare, în acest tip de învățare, rețeaua însăși trebuie să descopere tiparele, caracteristicile din datele de intrare și relația între datele de intrare si datele de ieșire.

Detecție bazată pe ansambluri

După cum știm, nici una dintre metodele de detectare anterioare nu poate descoperi toate anomaliile într-un sub-spațiu cu dimensiuni reduse datorită complexității datelor. Astfel, sunt necesare simultan diferite tehnici de învățare sau chiar mai multe sub spații, unde potențialele anomalii sunt derivate prin tehnici de ansamblu. Metoda de numita FB (Feature Bagging) vizează instruirea mai multor modele pe diferite subseturi de caracteristici și apoi combină rezultatele modelului într-o decizie generală.

Tehnicile bazate pe ansamblu pot atinge o precizie mai mare decât algoritmii simpli, dar principalul dezavantaj este că trebuie să instruim mai multe modele, care pot consuma mult timp.

Sisteme hibride

Cea mai recentă dezvoltare a tehnologiei de detecție a anomaliilor sunt sistemele hibride. Sistemele hibride încorporează algoritmi din cel puțin două dintre secțiunile precedente (metode statistice, neuronale sau de învățare automată). Hibridizarea este utilizată în mod diferit pentru a depăși deficiențele folosind un singur tip de algoritm, pentru a exploata avantajele abordărilor multiple, depășind în același timp punctele slabe ale acestora sau folosind un meta-clasificator pentru a agrega rezultatele de la mai mulți clasificatori pentru a gestiona toate situațiile.

Observații finale

Între metodele statistice tradiționale putem enumera metode care utilizează un prag simplu sau o fereastră glisantă (sliding window), metode care utilizează medie, mediană, deviație standard, distribuție etc. Există și algoritmi mai complecși care se bazează pe anumite statistici, cum ar fi detecția anomaliilor bazată pe histogramă, metoda liniară pentru detecția anomaliilor bazată pe deviația standard (LMDD - Deviation-based Outlier Detection) [35][14], Integrala de corelație locală (LOCI - Local Correlation Integral) [36][15], Determinantul de covarianță minimă (MCD - Minimum Covariance Determinant) [37][16] sau

Selecție stochastică a anomaliilor (SOS - Stochastic Outlier Selection). Problema care apare atunci când folosim metode statistice este că definiția normalului și a anormalului se poate schimba, astfel încât regulile statistice să nu facă față schimbărilor frecvente.

În cazul algoritmilor pe bază de densitate, presupunerea este că punctele de date normale apar în jurul unor vecinătăți dense. Algoritmii utilizați în acest caz sunt diferite variații ale tehnicii numite Local Outlier Factor.

Ideea generală din spatele abordării bazate pe distanță este de a judeca un punct bazat pe distanța sau distanțele față de vecinii săi. Presupunerea este că un obiect de date normal are o vecinătate densă, deci distanțele dintre vecini sunt mici, în timp ce în cazul unui punct anormal punctele de date sunt departe de vecinii lor. În acest caz, metricile de evaluare sunt diferite metrici de distanță, cum ar fi distanța Euclidiană, Manhattan sau Minkowski în general. Cei mai obișnuiți algoritmi sunt kNN sau Connectivity-Based Outlier Factor.

Detecția anomaliilor bazate pe clustere este un subiect mai larg și conține mai multe tehnici diferite:

- Cluster-Based Local Outlier Factor (CBLOF) [38]
- Isolation Forest [39]
- One Class SVM [40]
- Principal Component Analysis [24]
- K-means [41]
- Subspace Outlier Detection [42]
- XGBoost [43]
- etc

Cea mai flexibilă tehnică de detecție a anomaliilor este utilizarea rețelelor neuronale artificiale. În acest caz, nu avem nicio ipoteză prealabilă despre domeniul problemei, despre date, nu trebuie să definim regulile manual, definiția datelor normale și anormale sunt deduse automat din datele de intrare.

Avem mai multe tipuri de rețele neuronale care pot fi folosite pentru detecția anomaliilor:

- Rețele temporale spațiale (STN - Spatio Temporal Networks)[44][45][46][47]
- Rețele sumă de produse (SPN - Sum-Product Networks)[48]
- Modele generative [49][50][51][52][53]
- Rețele neuronale convoluționale[54][55]
- Rețele neuronale recurente [56]
- Auto-encoderi [57][58][59][60]

Bibliografie

[1] Portalul Natiunilor Unite https://www.un.org/en/sections/issues-depth/water/ accesat la 27.11.2020

[2] https://www.eea.europa.eu/publications/public-health-and-environmental-protection

[3] Directiva europeana 2000/60/CE https://eur-lex.europa.eu/legal-content/RO/LSU/?uri=celex:32000L0060 accesat la 26.11.2020

[4] Portalul Comisiei Europene, infografic referitor la apa https://ec.europa.eu/environment/water/infographics.htm accesat la 27.11.2020

[5] Portalul Ministerul Mediului si schimbarilor climatice http://www.mmediu.ro/beta/domenii/managementul-apelor-2/managementul-resurselor-de-apa/ accesat la 25.11.2020

[6] Guvernul Australiei de Vest, Ministerul Apei, https://www.water.wa.gov.au/__data/assets/pdf_file/0018/2934/87153.pdf accesat la 25.11.2020

[7] https://www.unwater.org/water-facts/

[8] Portal Comisia Europeana, proiectul WATMAN https://ec.europa.eu/regional_policy/en/projects/Romania/water-management-system-to-reduce-flooding-risks-in-romania accesat la 25.11.2020

[9] Portal LSI Lastem, https://www.lsi-lastem.com/PDF/MW9000/MW9000-ENG-01-Air-temperature.pdf accesat la 27.11.2020

[10] Portal Logotronic pt. senzori multiparametrici http://www.logotronic.at/wp-content/uploads/2018/01/Eureka_Manta2.pdf accesat la 28.11.2020

[11] Life Smart water: https://www.life-smartwater.com/technology

[12] K. Ferencz, J. Domokos, Overview of Industrial IoT services and open source systems, Conferinta SzamOkt 2020

[13] https://azure.microsoft.com/en-us/overview/iot/

[14] https://aws.amazon.com/iot/

[15] https://cloud.google.com/iot/

[16] OGC Sensor Observation Service Standard https://www.ogc.org/standards/sos

[17] OGC SensorThingsAPI https://www.ogc.org/standards/sensorthings

[18] https://projects.eclipse.org/proposals/whiskers

[19] "Geodan/gost". GitHub.

[20] "mozilla-sensorweb/sensorthings" GitHub.

[21] https://github.com/52North/sensorweb-server-sta

[22]„Numpy NDArrays”, The SciPy community, https://numpy.org/doc/stable/reference/generated/numpy.ndarray.html, 2020

[23]X. Chen and J. C. Jeong, "Enhanced recursive feature elimination," Sixth International Conference on Machine Learning and Applications (ICMLA 2007), Cincinnati, OH, 2007, pp. 429-435, doi: 10.1109/ICMLA.2007.35.

[24]Svante Wold, Kim Esbensen, Paul Geladi, Principal component analysis, Chemometrics and Intelligent Laboratory Systems, Volume 2, Issues 1–3, 1987, Pages 37-52, ISSN 0169-7439, https://doi.org/10.1016/0169-7439(87)80084-9.

[25]Z. Zhang, Y. Takane, Multidimensional Scaling, Editor(s): Penelope Peterson, Eva Baker, Barry McGaw, International Encyclopedia of Education (Third Edition), Elsevier, 2010, Pages 304-311, ISBN 9780080448947, https://doi.org/10.1016/B978-0-08-044894-7.01348-8.
[26]Dick de Ridder, Robert P.W. Duin, Sammon's mapping using neural networks: A comparison, Pattern Recognition Letters, Volume 18, Issues 11–13, 1997, Pages 1307-1316, ISSN 0167-8655, https://doi.org/10.1016/S0167-8655(97)00093-7.
[27]N. Rogovschi, J. Kitazono, N. Grozavu, T. Omori and S. Ozawa, "t-Distributed stochastic neighbor embedding spectral clustering," 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, 2017, pp. 1628-1632, doi: 10.1109/IJCNN.2017.7966046.
[28]Daniela Bianchi, Raffaele Calogero, Brunello Tirozzi, Kohonen neural networks and genetic classification, Mathematical and Computer Modelling, Volume 45, Issues 1–2, 2007, Pages 34-60, ISSN 0895-7177, https://doi.org/10.1016/j.mcm.2006.04.004.
[29]J. W. Sammon, "A Nonlinear Mapping for Data Structure Analysis," in IEEE Transactions on Computers, vol. C-18, no. 5, pp. 401-409, May 1969, doi: 10.1109/T-C.1969.222678.
[30]F. E. Grubbs, “Procedures for detecting outlying observations in samples,” Technometrics11, 1–21 (1969).
[31] F. Angiulli and C. Pizzuti, “Fast outlier detection in high dimensional spaces,” in Proceedingsof the 6th European Conference on Principles of Data Mining and Knowledge Discovery,(Springer-Verlag, Berlin, Heidelberg, 2002), PKDD ’02, p. 15–26.
[32]M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, “Lof: Identifying density-based localoutliers,” inProceedings of the 2000 ACM SIGMOD International Conference on Management ofData,(Association for Computing Machinery, New York, NY, USA, 2000), SIGMOD ’00, p.93–104
[33]J. Tang, Z. Chen, A. W.-c. Fu, and D. W. Cheung, “Enhancing effectiveness of outlier detec-tions for low density patterns,” inAdvances in Knowledge Discovery and Data Mining,M.-S.Chen, P. S. Yu, and B. Liu, eds. (Springer Berlin Heidelberg, Berlin, Heidelberg, 2002), pp.535–548
[34]H. Kriegel, P. Kröger, E. Schubert, and A. Zimek, “Loop: Local outlier probabilities,” inProceedings of the 18th ACM conference on Information and knowledge management,(Associationfor Computing Machinery, United States, 2009), pp. 1649–1652. ACM 18th InternationalConference on Information and Knowledge Management ; Conference date: 02-11-2009Through 06-11-2009
[35]Z. Zhang and X. Feng, "New Methods for Deviation-Based Outlier Detection in Large Database," 2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery, Tianjin, 2009, pp. 495-499, doi: 10.1109/FSKD.2009.303.
[36]S. Papadimitriou, H. Kitagawa, P. B. Gibbons and C. Faloutsos, "LOCI: fast outlier detection using the local correlation integral," Proceedings 19th International Conference on Data Engineering (Cat. No.03CH37405), Bangalore, India, 2003, pp. 315-326, doi: 10.1109/ICDE.2003.1260802.
[37]Hubert, Mia & Debruyne, Michiel & Rousseeuw, Peter. (2018). Minimum Covariance Determinant and Extensions. Wiley Interdisciplinary Reviews: Computational Statistics. 10. e:1421. 10.1002/wics.1421.
[38]Z. Gao, "Application of Cluster-Based Local Outlier Factor Algorithm in Anti-Money Laundering," 2009 International Conference on Management and Service Science, Wuhan, 2009, pp. 1-4, doi: 10.1109/ICMSS.2009.5302396.

[39]Liu, Fei Tony & Ting, Kai & Zhou, Zhi-Hua. (2009). Isolation Forest. 413 - 422. 10.1109/ICDM.2008.17.
[40]Amer, Mennatallah & Goldstein, Markus & Abdennadher, Slim. (2013). Enhancing one-class Support Vector Machines for unsupervised anomaly detection. Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description, ODD 2013. 8-15. 10.1145/2500853.2500857.
[41]Youguo Li, Haiyan Wu, A Clustering Method Based on K-Means Algorithm, Physics Procedia, Volume 25, 2012, Pages 1104-1109, ISSN 1875-3892, https://doi.org/10.1016/j.phpro.2012.03.206
[42]S. Sathe and C. C. Aggarwal, "Subspace Outlier Detection in Linear Time with Randomized Hashing," 2016 IEEE 16th International Conference on Data Mining (ICDM), Barcelona, 2016, pp. 459-468, doi: 10.1109/ICDM.2016.0057.
[43]Chen, Tianqi & Guestrin, Carlos. (2016). XGBoost: A Scalable Tree Boosting System. 785-794. 10.1145/2939672.2939785.
[44]S. Lee, H. Kim, and Y. Ro, “Stan: Spatio-temporal adversarial networks for abnormal eventdetection,” (2018).
[45]M. Szekér, “Spatio-temporal outlier detection in streaming trajectory data,” (2014).
[46]L. Nie, Y. Li, and X. Kong, “Spatio-temporal network traffic estimation and anomaly detec-tion based on convolutional neural network in vehicular ad-hoc networks,” IEEE Access6,40168–40176 (2018)
[47]E. W. Dereszynski and T. G. Dietterich, “Spatiotemporal models for data-anomaly detectionin dynamic environmental monitoring campaigns,” ACM Trans. Sen. Netw.8(2011).
[48]R. Peharz, A. Vergari, K. Stelzner, A. Molina, M. Trapp, K. Kersting, and Z. Ghahramani,“Probabilistic deep learning using random sum-product networks,” (2018)
[49]I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville,and Y. Bengio, “Generative adversarial networks,” (2014).
[50]D. Li, D. Chen, J. Goh, and S. kiong Ng, “Anomaly detection with generative adversarialnetworks for multivariate time series,” (2018).
[51]L. Deecke, R. Vandermeulen, L. Ruff, S. Mandt, and M. Kloft, “Anomaly detection withgenerative adversarial networks,” (2018).
[52]T. Schlegl, P. Seeböck, S. Waldstein, U. Schmidt-Erfurth, and G. Langs, “Unsupervised anomaly detection with generative adversarial networks to guide marker discovery,” (2017),pp. 146–157.
[53]M. Ravanbakhsh, M. Nabi, E. Sangineto, L. Marcenaro, C. Regazzoni, and N. Sebe, “Abnor-mal event detection in videos using generative adversarial nets,” (2017).
[54]A. Krizhevsky, I. Sutskever, and G. Hinton, “Imagenet classification with deep convolutionalneural networks,” Neural Inf. Process. Syst.25(2012).
[55]O. Gorokhov, M. Petrovskiy, and I. Mashechkin, “Convolutional neural networks for unsu-pervised anomaly detection in text data,” (2017), pp. 500–507.
[56]T. Ergen and S. S. Kozat, “Unsupervised anomaly detection with lstm neural networks,”IEEE Transactions on Neural Networks Learn. Syst. p. 1–15 (2019).
[57]S. Hawkins, H. He, G. Williams, and R. Baxter, “Outlier detection using replicator neuralnetworks,” (2002), pp. 113–123.
[58]D. Wulsin, J. Blanco, R. Mani, and B. Litt, “Semi-supervised anomaly detection for eegwaveforms using deep belief nets,” in2010 Ninth International Conference on Machine Learningand Applications,(2010), pp. 436–441.

[59]M. Nadeem, O. Marshall, S. Singh, X. Fang, and X. Yuan, “Semi-supervised deep neuralnetwork for network intrusion detection,” (2016).
[60]H. Song, Z. Jiang, A. Men, and B. Yang, “A hybrid semi-supervised anomaly detectionmodel for high-dimensional data,” Comput. Intell. Neurosci.2017, 1–9 (2017).
[61]Ye, Jieping & Janardan, Ravi & Li, Qi. (2004). Two-Dimensional Linear Discriminant Analysis.. Adv. Neural. Inf. Process. Syst.. 17.
[62]Mourtzios C. et al. (2021) Work-in-Progress: SMART-WATER, a Novel Telemetry and Remote Control System Infrastructure for the Management of Water Consumption in Thessaloniki. In: Auer M.E., Tsiatsos T. (eds) Internet of Things, Infrastructures and Mobile Applications. IMCL 2019. Advances in Intelligent Systems and Computing, vol 1192. Springer, Cham. https://doi.org/10.1007/978-3-030-49932-7_89
[63]S. Agrawal, J. Agrawal, Survey on Anomaly Detection using Data Mining Techniques, 19th International Conference on Knowledge Based and Intelligent Information and Engineering Systems, ed. Elsevier,Procedia Computer Science 60 (2015) 708 – 713
[64]M. Gupta, J. Gao, C. C. Aggarwal,J. Han, Outlier Detection for Temporal Data: A Survey, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 1, JANUARY 2014
[65]V. Barnett and T. Lewis, Outliers in Statistical Data, New York: John Wiley Sons, 1994.
[66]V. Chandola, A. Banerjee, V. Kumar, Anomaly Detection: A Survey, ACM Computing Surveys 41, 3 (2009).
[67]J. M. Estevez-Tapiador, P. Garcia-Teodoro, J. E. Diaz-Verdejo, Anomaly detection methods in wired networks: a survey and taxonomy, Computer Communications, 15. October 2014, volume 27
[68]Y. Zhang, N. Meratnia, P. Havinga, Outlier Detection Techniques for Wireless Sensor Networks: A Survey, IEEE COMMUNICATIONS SURVEYS AND TUTORIALS, VOL. 12, NO. 2, SECOND QUARTER 2010
[69]V.J. Hodge, J. Austin, A Survey of Outlier Detection Methodologies, Kluwer Academic Publishers, 2004