

Saulės elektrinių matavimų klasifikavimas remiantis NLP ir mašininio mokymusi

Lukas Voveris, Jolita Bernatavičienė

Vilniaus universitetas, Matematikos ir informatikos fakultetas,
Duomenų mokslo ir skaitmeninių technologijų institutas,
Akademijos g. 4, Vilnius
lukas.voveris@mif.stud.vu.lt, jolita.bernatavicene@mif.vu.lt

Santrauka. Sparčiai augant saulės elektrinių skaičiui, didėja ir poreikis efektyviai analizuoti jų veikimo duomenis. Visgi duomenų analizę dažnai apsunkina nevienodas matavimo kintamųjų žymėjimas, kylantis dėl skirtingų įrangos gamintojų, montuotojų ar lokalizacijos. Šiame darbe siūloma automatizuota kintamųjų klasifikavimo metodika, leidžianti standartizuoti žymas pasitelkiant natūralios kalbos apdorojimo technikas ir mašininio mokymosi algoritmus. Tyrime panaudoti duomenys iš 178 saulės elektrinių, o kiekvienas įrašas buvo susietas su viena iš 176 klasių pagal specialiai sukurtą septynių komponentų žymėjimo sistemą. Modelių mokymui buvo suformuoti skaitiniai požymiai, įskaitant TF-IDF reikšmes, matavimo vienetų kodavimą ir kontekstinius požymius, gautus kaimynystės analizės būdu. Rezultatai parodė, kad siūlomi modeliai pasiekia itin didelį klasifikavimo tikslumą (daugeliu atvejų virš 99,9 %), o geriausi rezultatai gauti naudojant sprendimų medžio ir atsitiktinių miškų metodus.

Raktiniai žodžiai: saulės elektrinės, klasifikavimas, natūralios kalbos apdorojimas, TF-IDF, mašininis mokymas, atraminių vektorių mašina, atsitiktiniai miškai, sprendimų medis, logistinė regresija.

1 Įvadas

Pastaraisiais metais visame pasaulyje saulės elektrinių diegimas sparčiai auga, skatinamas visuotinio perėjimo prie atsinaujinančios energijos ir būtinybės mažinti anglies dvideginio emisijas. 2023 metais pasaulinės saulės energijos instaliacijos augimas beveik padvigubėjo, naujai įdiegtos saulės elektrinės pridėjo 447 GW našumo į globalią rinką (palyginti 239 GW 2022 metais). 87 % augimas rodo spartų sektoriaus plėtimąsi ir vis didesnę saulės energijos reikšmę pasaulio energetikos sistemoje [1]. Tačiau, nepaisant didelio šių sistemų potencialo, jų veikimą gali trikdyti įvairūs nenumatyti veiksniai. Pavyzdžiui, saulės modulių efektyvumą mažina šėšėliavimas, dulkės,

netinkama modulio orientacija ir nepalankios oro sąlygos. Be to, aplinkos veiksniai, tokie kaip temperatūra, drėgmė ir vėjo greitis, irgi turi įtakos saulės modulių našumui [2]. Didėjant saulės fotovoltinių (PV) sistemų mastui ir sudėtingumui, tampa itin svarbu efektyviai stebėti jų darbą bei analizuoti surinktus duomenis, siekiant palaikyti našumą ir laiku aptikti gedimus [3]. Esminė šio proceso dalis – duomenų rinkimas iš įvairių komponentų, tokių kaip inverteriai, saulės moduliai ar meteorologinės stotelės. Tačiau bandant apjungti ar palyginti duomenis iš skirtingų elektrinių kyla rimtas iššūkis. Dėl įrangos gamintojų, monitoringo sistemų, įrangos montuotojų ar lokalizacijos skirtumų matavimų kintamųjų pavadinimai dažnai labai skiriasi. Tokie neatitikimai apsunkina duomenų standartizavimą ir tampa kliūtimi atlikti analizę, aptikti anomalijas ar rengti ataskaitas. Siekiant plėsti analizės modelius ar automatizuotas ataskaitas plačiu mastu, nevienodi pavadinimai ir žymėjimai trukdo automatiškai identifikuoti reikalingus kintamuosius. Dėl to būtina individualizuoti sprendimus beveik kiekvienai elektrinei, o tai apsunkina sprendimų taikymą platesniu mastu. Norint spręsti šią problemą, vis daugiau dėmesio skiriama unifikuotoms kintamųjų žymėjimo sistemoms kurti. Vienas iš tokių sprendimų yra SunSpec standartas, kuris apibrėžia bendrus parametrus ir nustatymus, skirtus stebėti ir valdyti paskirstytos energijos šaltinius [4]. Vis dėlto šis standartas orientuotas į „Modbus“ protokolą, todėl jo taikymas kitų tipų įrenginiams gali būti ribotas. Be to, SunSpec ir panašūs standartai pasižymi didžiausiu efektyvumu tuomet, kai jie yra įgyvendinami jau elektrinės projektavimo etapu, nes tai leidžia užtikrinti sistemų suderinamumą bei optimalų duomenų valdymą. Tačiau senesnėse elektrinėse, kurios buvo įdiegtos be šių standartų, būtina peržiūrėti ir pritaikyti esamas sistemas, siekiant suderinti kintamųjų pavadinimus su SunSpec ar kitais standartais. Tai reikalauja papildomų pastangų ir išteklių, nes kiekviena elektrinė gali turėti unikalių parametrų ir konfigūracijų, kuriuos reikia individualiai pritaikyti prie standartizuotos sistemos. Statistiniai metodai ir mašininio mokymosi modeliai gali reikšmingai prisidėti prie esamų matavimo pavadinimų standartizavimo proceso, leidžiant efektyviau atpažinti matavimų tipus pagal jų pavadinimus, matavimo vienetus ar prietaisų pavadinimus [5]. Tokios sistemos gali stipriai sumažinti rankinį darbą, padidinti duomenų integracijos mastą ir atverti kelią pažangesnei skirtingų saulės elektrinių analitikai. Šiame darbe pristatomas šios problemos sprendimas – siūlomas duomenų žymėjimo standartas bei metodas atpažinti ir klasifikuoti automatinio matavimo kintamuosius, kuris leidžia esamą žymėjimą transformuoti į naujai siūlomą standartą. Tyrimas inicijuotas UAB Elseta.

2 Duomenys

Tyrimui pasitelkti duomenys iš 178 saulės elektrinių. Kiekvienas įrašas atspindėjo vienos elektrinės specifinio komponento (pvz., inverterio, skaitiklio ar kt.) vieną matavimo kintamąjį. Bendrai buvo analizuojami 20 225 duomenų rinkinio įrašai. Kiekvienas įrašas turėjo penkis atributus: matavimo identifikacinį numerį, matavimo pavadinimą, šaltinio (prietaiso) pavadinimą, matavimo vienetus bei objekto identifikacinį numerį. Remiantis SunSpec standartizavimo principais [4], buvo sukurta personalizuota žymėjimo sistema. Kiekviena žyma sudaryta iš septynių dalių, kurios atskiriamos apatiniu brūkšniu (_). Jei tam tikra žymos dalis nėra taikoma konkrečiam matavimui, ji žymima brūkšneliu (-). Žymos struktūra yra nuosekli ir apima šiuos komponentus tokia eilės tvarka: įrenginio tipas, srovės tipas, matavimo tipas, matavimo eilės numeris, agregavimo metodas, matavimo periodas ir informacija apie importą ar eksportą. Tyrime klase laikomas kiekvienas unikalus žymos formatą atitinkantis septynių komponentų derinys. Kitaip tariant, kiekviena unikali žyma (pvz., „INV_String_1_1_--“) ar „INV_AC_1_1_--“) yra laikoma atskira klase. Pradinis duomenų rinkinys buvo sužymėtas rankiniu būdu, o galutiniame rinkinyje identifiкуotos 176 skirtingos klasės. Prieš taikant klasifikavimo algoritmus, tekstiniai duomenys buvo apdoroti pasitelkiant natūralios kalbos apdorojimo (NLP) technikas [6]. Tekstas buvo normalizuotas: paverstas mažosiomis raidėmis, pašalinti skyrybos ženklai bei specialieji simboliai. Iš tekstų pašalinti nereikšmingi žodžiai ir atliktas žodžių kamienų išskyrimas. Įrenginio pavadinimuose dažniausiai pasitaikančios įrenginių variacijos (pvz., „inverter“, „meter“) buvo suvienodintos į standartizuotas reikšmes (pvz., „INV“, „Meter“). Kadangi matavimo pavadinime matavimo numeriai, priklausomai nuo prietaiso, gali būti žymimi arba skaičiais, arba raidėmis, šios reikšmės buvo suvienodintos. Raidiniai žymėjimai pakeisti į skaitinius (A = 1, B = 2 ir t.t.), išskirtos ir užkoduotos visos skaitinės reikšmės, išsaugotos atskiru požymiu.

3 Metodologija

Surinkus ir sužymėjus duomenis pagal sukurta žymėjimo sistemą, buvo parengta metodologija, skirta automatizuotai klasifikuoti matavimus, taikant mašininio mokymosi metodus. Siekiant pagerinti modelių prognozavimo tikslumą, buvo sukurta papildomų požymių:

- Indikatoriai, rodantys, ar tekste yra tam tikrų reikšmių (pvz., „INV“, „DC“, „Current“, „Voltage“ ir pan.)

- Tekstiniai duomenys paversti į skaitinę formą naudojant TF-IDF [7] metodą.
- Matavimo vienetų reikšmės užkoduotos paverčiant skirtingas kategorines reikšmes į naujus požymius, kur 1 rodo atitinkamos reikšmės buvimą, o 0 – nebuvimą.

Šie požymiai buvo sujungti į bendrą požymių rinkinį, kuris naudotas visose tolimesnėse modelių mokymų fazėse.

Klasių prognozei, buvo pasirinkta naudoti keturis skirtingus klasifikavimo modelius: logistinę regresiją (LR) [8], sprendimų medį (SM) [9], atsitiktinių miškų (AM) [10] modelį, atraminių vektorių mašiną (SVM) [11]. Klasifikatoriai buvo vertinami taikant penkių dalių kryžminės validacijos metodą [12]. Gautos metrikos yra suskaičiuojamos kiekviename žingsnyje ir imamas jų vidurkis, kad būtų gautas bendras modelio našumo įvertis.

Modelių našumas buvo vertinamas pagal keturis pagrindinius klasifikacijos metrikų rodiklius: tikslumą, preciziškumą, jautrumą ir F1 įvertį. Šios metrikos buvo apskaičiuotos kiekvienai klasei atskirai, o bendra reikšmė gauta naudojant svertinį vidurkį, atsižvelgiant į kiekvienos klasės dydį.

Tikslumas apibūdina visų teisingai klasifikuotų įrašų santykį su bendru įrašų skaičiumi. Preciziškumas rodo, kiek modelio prognozuotų klasės i atvejų buvo teisingi. Jautrumas nurodo, kiek iš visų tikrųjų klasės i atvejų modelis atpažino teisingai. Tuo tarpu F1 įvertis – tai harmoninis preciziškumo ir jautrumo vidurkis, kuris suteikia subalansuotą vertinimą tarp abiejų.

$$Tikslumas = \frac{\sum_i^C (TP(i) + TN(i))}{\sum_i^C (TP(i) + TN(i) + FP(i) + FN(i))}$$

$$Preciziškumas(i) = \frac{TP(i)}{TP(i) + FP(i)}$$

$$Jautrumas(i) = \frac{TP(i)}{TP(i) + FN(i)}$$

$$F1(i) = \frac{2 \cdot Preciziškumas(i) \cdot Jautrumas(i)}{Preciziškumas(i) + Jautrumas(i)}$$

Čia, $TP(i)$ – kiek kartų klasė i buvo teisingai atpažinta kaip klasė i . $TN(i)$ – kiek kartų nei tikroji, nei prognozuota klasė nebuvo i . $FP(i)$ – kiek kartų kita klasė buvo klaidingai atpažinta kaip klasė i . $FN(i)$ – kiek kartų klasė i buvo klaidingai priskirta kitai klasei.

Norint gauti bendrą preciziškumo, jautrumo ir F1 įverčio reikšmę, buvo skaičiuojamas svertinis vidurkis, atsižvelgiant į kiekvienos klasės dydį:

$$Bendrasis\ ivertis = \frac{\sum_{i=1}^C s_i \cdot Metrika(i)}{\sum_{i=1}^C s_i}$$

Čia, s_i – kiek įrašų priklauso klasei i , $Metrika(i)$ – atitinkamos metrikos įvertis klasei i .

4 Eksperimentai ir rezultatai

Po duomenų žymėjimo buvo atlikta galimų klasių dažnių analizė, kuri parodė, kad kai kurioms klasėms (pvz. „INV_AC_Q_--_--“) priklauso vos vienas stebėjimas. Tai gali apsunkinti šių klasių patikimą identifikavimą. Apdorojus duomenis, gautas duomenų rinkinys, sudarytas iš 20225 eilučių ir 1819 stulpelių. Modelių vertinimui buvo taikyta penkių dalių kryžminė validacija. Kiekviename validacijos cikle keturios dalys buvo naudojamos modelio mokymui, o likusi – testavimui. Klasifikatorių – LR, AM, SM bei SVM – našumas buvo vertinamas apskaičiuojant vidutines metrikas iš visų penkių validacijos kartojimų.

Iš rezultatų matyti, kad aukščiausią tikslumą pasiekė SVM klasifikatorius, kurio tikslumas siekė 92,56 %. Vis dėlto toks rezultatas nėra idealus. Atlikus prognozuotų ir realių reikšmių analizę, nustatyta, kad didžioji dalis klaidų buvo susijusios su matavimo eilės numerio neatitikimais. Pavyzdžiui, modelis įrašą klasifikavo kaip „INV_String_I_7_--_--“ – septintos inverterio grandinės srovę, kai tikroji reikšmė buvo „INV_String_I_6_--_--“ – šeštos inverterio grandinės srovė. Tokie neatitikimai rodo, kad klasifikatoriui sunku tiksliai identifikuoti aukštesnės eilės grandinės matavimus, kurių stebėjimų skaičius duomenyse buvo nedidelis. Remiantis pirminiais rezultatais, nuspręsta klasifikavimo užduotį suskaidyti į kelias atskiras prognozavimo užduotis, siekiant padidinti tikslumą ir palengvinti interpretaciją. Vietoje vienos sudėtinės klasės buvo suformuotos septynios atskiros kategorijos: įrenginio tipas („Device“), srovės tipas („Current“), matavimo tipas („Type“), matavimo eilės numeris („InputId“), agregavimo metodas („Aggregation“), matavimo periodas („Period“) ir informacija apie importą ar eksportą („Import/Export“). Duomenų rinkinio analizėje kiekviena kategorija buvo atskirai prognozuojama taikant keturis skirtingus klasifikatorius – LR, SM, AM ir SVM. Modelių našumo vertinimui buvo pasitelktas penkių dalių kryžminės validacijos metodas. Galutiniai metrių įverčiai buvo apskaičiuoti kaip visų

validacijos ciklų vidurkiai. Visų klasifikatorių mokymui buvo taikytas vienas požymių rinkinys, tačiau kiekvienos kategorijos atveju, modeliai buvo apmokomi prognozuoti tos kategorijos klases.

Iš rezultatų matyti, kad visų septynių kategorijų klasifikatorių našumas yra ypač aukštas – beveik visais atvejais tikslumas, preciziškumas, jautrumas ir F1 rodikliai viršija 99 %. Vienintelio matavimo eilės numerio klasifikavimas geriausiai atliekamas naudojant SM klasifikatorių, kuris pasiekė 92,58 % tikslumą. Atliekant klaidų analizę, paaiškėjo, kad dažniausiai klaidų kyla, kai tam pačiam įrenginiui priskiriami keli vienodo tipo matavimai su tapačiu matavimo eilės numeriu. Tokiais atvejais modelis sunkiai atskiria panašius įrašus. Kitas dažnas netikslumų šaltinis – fazių matavimai, kurie turėtų būti žymimi kaip A-B, B-C, C-A (atitinkamai 1-2, 2-3, 3-1). Visgi klasifikavimo metu jie neretai interpretuojami tik pagal pirmąją fazę, t. y. 1, 2 arba 3.

Siekiant padidinti tikslumą ir išspręsti šias problemas, buvo nuspręsta atlikti kelis duomenų apdorojimo etapo pakeitimus. Pirmiausia, frazių grupių nominalūs žymėjimai buvo pakeisti skaitinėmis išraiškomis. Siekiant aiškiau atskirti fazių numerius nuo įprastų numerių žymėjimo, buvo atlikti tokie pakeitimai:

- A-B → 100-200
- B-C → 200-300
- C-A → 300-100

Šios reikšmės pasirinktos tam, kad fazių žymėjimai būtų aiškiai atskirti nuo kitų numeravimo reikšmių. Bandant spręsti vienodų eilės numerių vienodo matavimo tame pačiame įrenginyje problemą, buvo atlikta kaimynystės analizė. Kiekvienam įrašui priskirti panašūs įrašai (kaimynai), atsižvelgiant į:

- Tą patį objekto identifikacinį numerį
- Sutampantį prietaiso pavadinimą
- Tokį patį matavimo pavadinimą (be skaičių)

Tokiu būdu buvo siekiama identifikuoti įrašus, kurie turi tą pačią struktūrą ar šabloną, tačiau skiriasi tik skaitinėmis reikšmėmis (pvz. įtampos fazėse ar grandinėse). Randami kiekvieno įrašo kaimynai ir išgaunamos jų skaitinės reikšmės. Šios reikšmės yra užkoduojamos ir paverčiamos į vieningą reprezentaciją – sukuriama binariniai požymiai, tokie kaip neighbor_1,

neighbor_3, kuriuose reikšmė 1 reiškia, kad įrašas turi kaimyną su atitinkamu skaičiumi.

Įrašui priskiriamas dvejetainis požymis, kuris įgyja reikšmę 1, jei bent vienas kaimynas turi tą pačią skaitinę reikšmę kaip ir pats įrašas, ir 0 – jei tokių sutapimų nėra. Papildomai apskaičiuojamas ir kiekybinis požymis, rodantis, kiek kitų įrašų toje pačioje kaimynystėje turi identišką skaitinę reikšmę. Šios papildomos savybės leidžia klasifikatoriams geriau įvertinti ne tik kiekvieno įrašo turinį atskirai, bet ir jo kontekstą – tai yra, kaip jis susijęs su kitais panašiais to paties įrenginio ar sistemos įrašais.

Atlikus šiuos duomenų apdorojimo patobulinimus, modeliai dar kartą buvo apmokyti, naudojant tuos pačius mokymo duomenis, ir įvertinti tais pačiais metodais bei metrikomis. Rezultatai pateikiami 1 lentelėje. Iš jų matyti, kad matavimų eilės numerių klasifikavimo tikslumas reikšmingai padidėjo – nuo 92,58 % iki 99,76 %.

Geriausius rezultatus gauti naudojant SM, AM ir SVM klasifikatorius. Prietaiso tipas buvo prognozuotas 99,99±0,01 % tikslumu, srovės tipo prognozės AM modeliu pasiektos 99,93±0,04 % tikslumu. Tipas – taip pat AM modeliu – 99,92±0,05 %. Matavimo numeris, agregavimo lygis ir importo ar eksporto požymis geriausiai prognozuoti SM modeliu, atitinkamai 99,76±0,09 %, 99,97±0,04 %, 99,99±0,01 % tikslumu. Matavimo periodas – SM, AM ir SVM modeliais – taip pat 99,95±0,02 % tikslumu.

Geriausiams kiekvienos kategorijos modeliams buvo atliktas hiperparametrų tyrimas, naudojant gardelinės paieškos metodą. SM modeliui naudota ši hiperparametrų gardelė: *max_depth* = [None, 10, 20, 30], *min_samples_split* = [2, 5, 10], *min_samples_leaf* = [1, 2, 4]. AM modeliui – *n_estimators* = [100, 200], *max_depth* = [None, 10, 20], *min_samples_split* = [2, 5], *min_samples_leaf* = [1, 2]. Prietaiso tipo, srovės tipo ir periodo klasifikavimui optimalus buvo AM modelis su parametrais: *max_depth*=None, *min_samples_leaf*=1, *min_samples_split*=2, *n_estimators*=100. Matavimo tipo klasifikavimui taip pat geriausiai pasirodė AM modelis, tačiau su *min_samples_split*=5. Matavimo eilės numeriui, agregacijos tipui ir žymai apie importą/eksportą tiksliausiai klasifikuoti pavyko SM modeliams su parametrais: *max_depth*=None, *min_samples_leaf*=1, *min_samples_split*=2.

1 lentelė. Klasifikavimo rezultatai skirtingoms kategorijoms po patobulinimo.

Kategorija	Metodas	Tikslumas	Preciziškumas	Jautrumas	F1
Device	LR	0,9992 ± 0,0001	0,9992 ± 0,0001	0,9992 ± 0,0001	0,9992 ± 0,0001
	SM	0,9999 ± 0,0002	0,9999 ± 0,0002	0,9999 ± 0,0002	0,9999 ± 0,0002
	AM	0,9999 ± 0,0002	0,9999 ± 0,0002	0,9999 ± 0,0002	0,9999 ± 0,0002
	SVM	0,9999 ± 0,0002	0,9999 ± 0,0002	0,9999 ± 0,0002	0,9999 ± 0,0002
Current	LR	0,9906 ± 0,0005	0,9909 ± 0,0004	0,9906 ± 0,0005	0,9906 ± 0,0005
	SM	0,9983 ± 0,0010	0,9983 ± 0,0010	0,9983 ± 0,0010	0,9983 ± 0,0010
	AM	0,9993 ± 0,0004	0,9993 ± 0,0004	0,9993 ± 0,0004	0,9993 ± 0,0004
	SVM	0,9918 ± 0,0012	0,9922 ± 0,0011	0,9918 ± 0,0012	0,9918 ± 0,0012
Type	LR	0,9979 ± 0,0006	0,9978 ± 0,0005	0,9979 ± 0,0006	0,9987 ± 0,0003
	SM	0,9987 ± 0,003	0,9987 ± 0,0003	0,9987 ± 0,0003	0,9987 ± 0,0003
	AM	0,9992 ± 0,0005	0,9992 ± 0,0005	0,9992 ± 0,0005	0,9992 ± 0,0005
	SVM	0,9991 ± 0,0003	0,9991 ± 0,0003	0,9991 ± 0,0003	0,9991 ± 0,0003
InputId	LR	0,9764 ± 0,0027	0,9777 ± 0,0028	0,9764 ± 0,0027	0,9760 ± 0,0026
	SM	0,9976 ± 0,0009	0,9976 ± 0,0009	0,9976 ± 0,0009	0,9976 ± 0,0009
	AM	0,9970 ± 0,0011	0,9970 ± 0,0011	0,9970 ± 0,0011	0,9970 ± 0,0011
	SVM	0,9915 ± 0,0016	0,9917 ± 0,0015	0,9915 ± 0,0016	0,9915 ± 0,0016
Aggregation	LR	0,9993 ± 0,0005	0,9993 ± 0,0005	0,9993 ± 0,0005	0,9993 ± 0,0005
	SM	0,9997 ± 0,0004	0,9997 ± 0,0004	0,9997 ± 0,0004	0,9997 ± 0,0004
	AM	0,9994 ± 0,0005	0,9994 ± 0,0005	0,9994 ± 0,0005	0,9994 ± 0,0005
	SVM	0,9996 ± 0,0004	0,9996 ± 0,0004	0,9996 ± 0,0004	0,9996 ± 0,0004
Period	LR	0,9998 ± 0,0002	0,9998 ± 0,0002	0,9998 ± 0,0002	0,9998 ± 0,0002
	SM	0,9999 ± 0,0002	0,9999 ± 0,0002	0,9999 ± 0,0002	0,9999 ± 0,0002
	AM	0,9999 ± 0,0002	0,9999 ± 0,0002	0,9999 ± 0,0002	0,9999 ± 0,0002
	SVM	0,9999 ± 0,0002	0,9999 ± 0,0002	0,9999 ± 0,0002	0,9999 ± 0,0002
Import/Export	LR	0,9994 ± 0,0002	0,9990 ± 0,0004	0,9994 ± 0,0002	0,9992 ± 0,0003
	SM	0,9999 ± 0,0001	0,9999 ± 0,0001	0,9999 ± 0,0001	0,9999 ± 0,0001
	AM	0,9997 ± 0,002	0,9996 ± 0,0004	0,9997 ± 0,0002	0,9996 ± 0,0003
	SVM	0,9999 ± 0,0002	0,9998 ± 0,0003	0,9999 ± 0,0002	0,9998 ± 0,0003

5 Išvados

Šiame darbe pristatyta metodologija, skirta automatizuotam saulės elektrinių matavimo kintamųjų klasifikavimui, pasitelkiant NLP technikas bei mašininio mokymosi algoritmus. Tyrimo rezultatai atskleidė, kad, nepaisant didelės duomenų įvairovės ir nevienodo skirtingų elektrinių žymėjimo, tin-

kamai parinkus duomenų paruošimo strategijas ir požymių inžinerijos metodus, galima pasiekti ypač aukštą klasifikavimo tikslumą.

Iš pradžių modeliai buvo mokomi klasifikuoti visas žymas kaip vientisus vienetus, tačiau tai lėmė klaidas specifinėse vietose, ypač prognozuojant matavimo eilės numerius. Siekiant padidinti prognozių tikslumą ir pagerinti interpretaciją, klasifikavimo užduotis buvo suskaidyta į septynias atskiras kategorijas, kurios atitiko žymų struktūrą. Toks sprendimas leido kiekvieną žymos komponentą prognozuoti atskirai, ženkliai padidinant bendrą tikslumą ir lankstumą.

Be to, atlikti papildomi duomenų paruošimo patobulinimai – fazių žymėjimo standartizavimas bei kaimynystės analizės pagrindu sukurti papildomi požymiai – dar labiau padidino prognozių tikslumą. Po šių patobulinimų klasifikavimo tikslumas sudėtingiausioje kategorijoje – matavimo eilės numeryje – išaugo nuo 92,58 % iki 99,76 %, o kitos kategorijos pasiekė virš 99,9% tikslumą.

Šie rezultatai rodo, kad automatizuotas matavimo kintamųjų klasifikavimas, pagrįstas NLP technikomis ir mašininio mokymosi modeliais, gali būti veiksmingai taikomas realiose situacijose, ypač integruojant nevienodai sužymėtus duomenis iš skirtingų saulės elektrinių. Tokie sprendimai gali stipriai sumažinti rankinį darbą, pagreitinti analizės procesus ir sudaryti sąlygas pažangesnėms stebėsenos bei ataskaitų sistemoms saulės energijos sektoriuje.

Padėka: Tyrimas finansuojamas pagal LR Švietimo, mokslo ir sporto ministerijos programą „Universitetų ekselencijos iniciatyvos“ (LR ŠMSM mokslo plėtros programos pažangos priemonė Nr. 12-001-01-01-01 „Gerinti mokslo ir studijų aplinką“).

Literatūra

- [1] Walburga Hemetsberger, Michael Schmela, Sonia Dunlop. SolarPower Europe (2024): Global Market Outlook for Solar Power 2024-2028.
- [2] Vidyandandan, K. V. (2017). An overview of factors affecting the performance of solar PV systems. *Energy Scan*, 27(28), 216.
- [3] Avishek, M., Zahnd, A., McKay, K. H., & Ellul, A. (2009). The importance of monitoring and performance: Analysis of a rural solar PV electrification project. In *Solar09, the 47th ANZSES Annual Conference*.
- [4] SolarEdge Inverters - SunSpec Logging, Technical Note. Version 3.0, March 2025. <https://knowledge-center.solaredge.com/sites/kc/files/sunspec-implementation-technical-note.pdf>

- [5] Nicchiotti, G., Fromageat, L., & Etienne, L. (2016). Machine Learning Strategy for Fault Classification Using Only Nominal Data. PHM Society European Conference, 3(1). <https://doi.org/10.36001/phme.2016.v3i1.1581>
- [6] Leila Rouka, Cemal Köse, and Parham Tofighi, „Using Machine Translators in Textual Data Classification,” *International Journal of Computer and Communication Engineering* vol. 1, no. 1, pp. 43-46 , 2012.
- [7] Fauzi, M., Arifin, A., & Yuniarti, A. (2017). Arabic book retrieval using class and book index based term weighting. *International Journal of Electrical and Computer Engineering (Ijece)*, 7(6), 3705. <https://doi.org/10.11591/ijece.v7i6.pp3705-3710>
- [8] LaValley, M. P. (2008). Logistic regression. *Circulation*, 117(18), 2395-2399.
- [9] De Ville, B. (2013). Decision trees. *Wiley Interdisciplinary Reviews: Computational Statistics*, 5(6), 448-455.
- [10] Breiman, L. (2001). Random forests. *Machine learning*, 45, 5-32.
- [11] Suthaharan, S., & Suthaharan, S. (2016). Support vector machine. *Machine learning models and algorithms for big data classification: thinking with examples for effective learning*, 207-235.
- [12] Berrar, D. (2019). Cross-validation.