UNIVERSITATEA BABEŞ-BOLYAI
FACULTATEA DE MATEMATICĂ ȘI INFORMATICĂ

Modele de învățare automată pentru prognoza pe termen scurt a vremii

Rezumatul tezei de doctorat

Student doctorand: Mihai Andrei
Conducător științific: Prof. dr. Czibula Gabriela

2021

Cuvinte cheie: Prognoza pe termen scurt a vremii, învățare automată, învățare profundă, reguli de asociere relaţională, reţele neuronale profunde, reţele cu autoorganizare.

Cuprins

Cuprinsul tezei de doctorat **2**

Lista publicațiilor **5**

Introducere **7**

1 Bază teoretică **13**
1.1 Prognoza pe pe termen scurt a vremii 13
1.2 Modele de învățare automată 14

2 Noi modele de învățare nesupervizată pentru analiza datelor meteorologice **16**
2.1 Setul de date radar 17
2.2 Analiza schimbării datelor radar în timp 18
2.3 Analiza șabloanelor din tranzițiile datelor dintre scanări consecutive ale radarului . . 19
2.4 Extinderea analizei pentru mai multe produse radar 20

3 Contribuții în dezvoltarea de modele de învățare profundă pentru prognoza pe termen scurt a vremii **22**
3.1 NowDeepN: o abordare pentru prognoza pe pe termen scurt a vremii folosind rețele neuronale profunde 23
3.2 $RadRAR$: O abordare bazată pe reguli de asociere relațională pentru predicția pe pe termen scurt a datelor radar 25
3.3 $XNow$: O tehnică de învățare profundă convoluțională pentru prognoza pe termen scurt a vremii bazată pe date radar 26

Concluzii **29**

Bibliografie **30**

Cuprinsul tezei de doctorat

List of Figures **3**

Glossary **6**

List of Tables **7**

List of publications **9**

Introduction **11**

1 Background **18**
1.1 Weather nowcasting . . . 19
1.1.1 Radar data . . . 19
1.1.2 Literature review . . . 20
1.1.2.1 Existing methods for weather nowcasting . . . 20
1.1.2.2 Multi-agent Systems for Weather Forecasting . . . 22
1.1.2.3 Supervised learning based approaches for weather nowcasting . . . 24
1.1.2.4 Unsupervised learning based approaches for meteorological data analysis . . . 30
1.2 Machine learning models . . . 31
1.2.1 Unsupervised learning . . . 32
1.2.2 Relational association rule mining . . . 32
1.2.3 Neural networks and deep learning . . . 33
1.2.4 Recurrent neural networks . . . 34
1.2.5 Convolutional neural networks . . . 36
1.2.6 Convolutional Long-short term memory networks (convLSTM) . . . 37

2 New unsupervised learning models for meteorological data analysis **38**
2.1 Radar data set . . . 39
2.2 Analysis of radar data change over multiple timestamps . . . 40
2.2.1 Methodology . . . 42
2.2.1.1 The proposed data model . . . 42
2.2.1.2 Experiment . . . 43
2.2.2 Experimental evaluation . . . 43
2.2.2.1 Statistical data analysis . . . 43
2.2.2.2 Experimental setup . . . 45
2.2.3 Results and discussion . . . 45
2.2.3.1 Analysis of the results from a meteorological perspective . . . 47

2.2.4 Conclusions and further work . . . 47
2.3 Analysis of patterns in radar data transition between consecutive radar scans . . . 48
2.3.1 Methodology . . . 49
2.3.1.1 Data model . . . 49
2.3.1.2 Experiments . . . 50
2.3.2 Results . . . 51
2.3.2.1 Statistical Analysis . . . 51
2.3.2.2 Self-Organizing Map (SOM) Results . . . 52
2.3.2.3 Discussion . . . 54
2.3.3 Conclusions and further work . . . 55
2.4 Extension of analysis for multiple radar products . . . 55
2.4.1 Methodology . . . 56
2.4.2 Experimental results and discussion . . . 58
2.4.3 Conclusions and future work . . . 61

3 Contributions in developing deep learning models for weather nowcasting 63
3.1 NowDeepN: An approach for nowcasting prediction using deep neural networks . . . 64
3.1.1 Methodology . . . 65
3.1.1.1 Data model . . . 65
3.1.1.2 Data collection and cleaning . . . 67
3.1.1.3 Building the $NowDeepN$ model . . . 69
3.1.1.4 Testing . . . 69
3.1.2 Experimental results . . . 70
3.1.2.1 Data set . . . 70
3.1.2.2 Data analysis . . . 71
3.1.2.3 Results . . . 72
3.1.3 Discussion . . . 76
3.1.3.1 Impact of the data cleaning step . . . 76
3.1.3.2 Relevance of the used features . . . 78
3.1.3.3 Comparison to related work . . . 79
3.1.4 Conclusions and future work . . . 81
3.2 $RadRAR$: A relational association rule mining approach for nowcasting based on predicting radar products' values . . . 82
3.2.1 Methodology . . . 83
3.2.1.1 Data model . . . 83
3.2.1.2 $RadRAR$ classifier . . . 84
Classification using $RadRAR$. . . . 85
3.2.1.3 Testing . . . 86
3.2.2 Results and discussion . . . 86
3.2.2.1 Data sets . . . 86
3.2.2.2 RAR mining . . . 87
3.2.2.3 Results and comparison to related work . . . 88
3.2.3 Conclusions and further work . . . 90
3.3 $XNow$: A convolutional deep learning technique for nowcasting based on radar products' values prediction . . . 90
3.3.1 Methodology . . . 91
3.3.1.1 Data model and preprocessing . . . 91

3.3.1.2 Xception deep learning model 92
3.3.1.3 Building the $XNow$ model 92
3.3.1.4 Evaluation 94
3.3.2 Results and discussion 94
3.3.2.1 Data set 95
3.3.2.2 Results 95
3.3.2.3 Discussion and comparison to related work 97
3.3.3 Conclusions and future work 97

Conclusions **99**

References **102**

Lista publicațiilor

Clasamentul publicațiilor a fost realizat conform standardelor CNATDCU (Consiliul Naţional de Atestare a Titlurilor, Diplomelor si Certificatelor Universitare) aplicabile pentru studenții doctoranzi înscriși după 1 octombrie 2018. Toate clasamentele sunt listate conform clasificării jurnalelor [1] și a conferințelor [2] în Informatică.

Publicații indexate în Web of Science - Science Citation Index Expanded

[CMA+21] Gabriela Czibula, **Andrei Mihai**, Alexandra-Ioana Albu, Istvan Czibula, Sorin Burcea, Abdelkader Mezghani. *AutoNowP: An approach using deep autoencoders for precipitation nowcasting based on radar echo prediction*. Mathematics, Special Issue on "Computational Optimizations for Machine Learning". 2021; in press. (**2020 IF=2.258**).

Rank A, 2 points.

[CMt21] Gabriela Czibula, **Andrei Mihai**, Eugen Mihuleţ. *NowDeepN: An ensemble of deep learning models for weather nowcasting based on radar products' values prediction*. Applied Sciences, Special Issue on"Applied Machine Learning". 2021; 11(1):125. (**2020 IF=2.679**).

Rank B, 4 points.

PPublicații indexate în Web of Science, Conference Proceedings Citation Index

[CMC19b] Gabriela Czibula, **Andrei Mihai**, Istvan G. Czibula, *RadRAR: A relational association rule mining approach for nowcasting based on predicting radar products' values.* 24th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES2020), Procedia Computer Science, Volume 176, 2020, pp. 300-309.

Rank B, 4 points.

[CMtT19] Gabriela Czibula, **Andrei Mihai**, Eugen Mihuleţ and Daniel Teodorovici. *Using self-organizing maps for unsupervised analysis of radar data for nowcasting purposes*. 23rd International Conference on Knowledge-Based and Intelligent Information & Engineering Systems, 2019, Procedia Computer Science Vol 159, (2019) pp. 48–57.

Rank B, 2 points.

[1] https://uefiscdi.ro/premierea-rezultatelor-cercetarii-articole
[2] http://portal.core.edu.au/conf-ranks/ Source CORE 2018

[CMC19a] Gabriela Czibula, **Andrei Mihai**, Liana Maria Crivei. *SPRAR: A novel relational association rule mining classification model applied for academic performance prediction.* International Conference on Knowledge Based and Intelligent Information and Engineering Systems (KES), 2019, Procedia Computer Science Vol 159, (2019) pp. 20–29

Rank B, 4 points.

[CAG19] Liana Maria Crivei, **Andrei Mihai**, Gabriela Czibula. *A study on applying relational association rule mining based classification for predicting the academic performance of students.* The 12th International Conference on Knowledge Science, Engineering and Management (KSEM), LNAI 11775, 2019, pp. 287-300.

Rank B, 4 points.

[MCt19] **Andrei Mihai**, Gabriela Czibula and Eugen Mihuleţ. *Analyzing Meteorological Data Using Unsupervised Learning Techniques*. 2019 IEEE 15th International Conference on Intelligent Computer Communication and Processing (ICCP), IEEE Computer Society Press, pp. 529 — 536. (**ISI Proceedings**).

Rank C, 2 points.

[Mih20] **Andrei Mihai**. *Using self-organizing maps as unsupervised learning models for meteorological data mining.* IEEE 13th International Symposium on Applied Computational Intelligence and Informatics, SACI 2020, Timișoara, pp. 23–28

Rank C, 2 points.

[SCIM20] Ioana Angela Socaci, Gabriela Czibula, Vlad-Sebastian Ionescu, **Andrei Mihai**. *A deep learning technique for nowcasting based on radar products' values prediction.* IEEE 13th International Symposium on Applied Computational Intelligence and Informatics, SACI 2020, Timișoara, pp. 117–122

Rank C, 1 point.

Scorul publicațiilor: 25 points.

Introducere

Domeniul de cercetare al tezei de doctorat este *învățarea profundă (deep learning)* aplicată în domeniul meteorologiei. Teza noastră de doctorat este intitulată 'Machine learning models for weather nowcasting" („Modele de învățare automată pentru prognoza pe pe termen scurt a vremii") și are scopul de a dezvolta noi modele de învățare automată pentru îmbunătățirea prognozei pe termen scurt a vremii (*nowcasting*).

Predicția vremii și, în special, a vremii severe, este o provocare importantă atât pentru cercetătorii din domeniul meteorologiei, cât și pentru cei din domeniul învățării automate. Complexitatea și dificultatea problemei se datorează în principal caracterului haotic al atmosferei, cât și setului implicit de informații meteorologice (radar, satelit sau observații meteorologice terestre) care trebuie analizate de meteorologi. Astfel, înțelegerea relațiilor dintre diferiți parametri meteorologici extrași din observațiile făcute de radar radar poate fi utilă pentru a oferi o înțelegere mai bună asupra evoluției vremii severe și ar ajuta la identificarea situațiilor în care poate apărea vreme severă.

Problema emiterii unei avertizări de tip *nowcasting* poate fi foarte dificilă pentru meteorologi, deoarece există adesea un set extrem de mare de date meteorologice (disponibile sub formă de observații radar, satelit sau observații meteorologice terestre) care trebuie analizate într-o perioadă foarte scurtă de timp. Prin urmare, metodele bazate pe *învățare automată* (în special cele *supervizate*) sunt necesare pentru obținerea unor soluții eficiente pentru problema nowcasting-ului. În plus, metodele de învățare *nesupervizate* sunt utile pentru extragerea de șabloane exacte și semnificative din cantitatea mare de date legate de vreme și pentru îmbunătățirea luării deciziilor în caz de vreme cu impact ridicat.

În primul rând, pentru a înțelege mai bine datele, am folosit *învățarea nesupervizată*. *Învățara nesupervizată* este un sub-domeniu al învățării automate care tratează algoritmi care extrag informații utile din date brute fără a utiliza exemple etichetate (cum se face în învățarea supervizată). De exemplu, metodele de învățare nesupervizate ar putea organiza datele în anumite clustere bazate pe o anumită funcție similaritate sau ar putea furniza un codificator și un decodor cu care datele de un anumit tip să fie comprimate și decomprimate sau poate extragerea unor reguli din date, fiind dată o structură pentru reguli. Pentru aceasta ne-am bazat pe *rețele cu auto-organizare* (Self-Organizing Maps – SOMs) [SK99] un tip de rețele neuronale artificiale nesupervizate.

Apoi, pe baza informațiilor extrase, am creat modele de *învățare supervizată*, pentru predicția datelor meteorologice. *Învățarea supervizată* este un sub-domeniu al învățării automate, care se ocupă de ideea de a aproxima o mapare de la un anumit domeniu de date de intrare la un anumit domeniu de date de ieșire bazat pe exemple de perechi intrare-ieșire. Un *algoritm de învățare supervizat* generalizează datele de antrenament, producând o funcție care, dată fiind o dată de intrare, poate returna o aproximare suficient de apropiată a rezultatului corect. Pentru a crea modele de învățare supervizate, ne-am concentrat în principal pe modelele de *rețele neuronale*. *Rețelele neuronale* au fost modelate pentru a fi similare cu rețelele complexe de neuroni. Această morfologie a fost adoptată în informatică, prin construirea de sisteme dens interconectate care au ca elemente de bază niște unități care au ca date de intrare o serie de numere cu valoare reală și produc ca și ieșire o singură valoare reală

[Mit97]. De asemenea, am experimentat crearea de modele de învățare supervizată folosind *Reguli de asociere relațională (Relational Association Rules)*. *Regulile de asociere relațională* reprezintă o tehnică importantă de analiză și investigare a datelor, utilă în diferite sarcini de învățare automată, deoarece sunt capabile să exprime diferite tipuri de relații non-ordinale între atributele datelor.

Problema abordată

Principala problemă abordată în lucrarea noastră este *prognoza pe pe termen scurt* în domeniul meteorologiei – *nowcasting*. Termenul „nowcasting" este derivat din contracția cuvintelor „now forecast" ("acum" și "predicție"), menită să însemne predicția evenimentelor pe termen foarte scurt. *Predicția vremii de tip nowcasting* este analiza și predicția pe termen scurt a șabloanelor vremii, în general pentru următoarele de la 0 până la 6 ore, și prezintă un interes major în cadrul cercetării meteorologice.

După cum afirmă Organizația Meteorologică Mondială (OMM) vremea, și în special vremea severă, provoacă multe dezastre naturale și este responsabilă pentru multe daune și pierderi de vieți omenești. Întrucât numărul și intensitatea evenimentelor meteorologice severe crește în multiple regiuni ale lumii, problema predicției unor astfel de fenomene și emiterea de alerte meteo este în prezent unul dintre cele mai populare subiecte din meteorologie. Prognoza precisă pe pe termen scurt a vremii este un element cheie pentru emiterea de alerte meteo relevante.

În timp ce predicția numerică a vremii poate fi folosită cu succes pentru prognozele meteo generale, pentru predicții pe termen foarte scurt – nowcasting – nu este la fel de utilizabilă, deoarece folosește simulări exacte ale ecuațiilor fizice care guvernează modelul atmosferic, astfel având nevoie de mult timp și o mare puterea de calcul pentru a face predicții [TSM21]. Din acest motiv, majoritatea sistemelor de nowcasting bazate pe stadiul actual al tehnologiei utilizează alte metode – și anume extrapolarea datelor meteorologice [SCW+15]. Acest lucru evidențiază un aspect care face ca nowcasting-ul vremii să fie un subiect atât de complex: predicțiile trebuie să fie rapide, astfel încât alertele meteo să poată fi emise cât mai curând posibil.

Alt aspect care face ca prognoza pe pe termen scurt a vremii să fie o sarcină atât de complexă este cantitatea mare de date disponibile, care trebuie analizate pentru a face predicții nowcasting bune. În primul rând, există multe surse de date, și toate ar putea fi relevante. Există mulți sateliți de meteorologie care generează în continuu date, care adună date despre elemente cum ar fi temperatură, vânt sau nori; în timp ce pe pământ există stații terestre care adună în mod constant date în timp real, de la stații radar la stațiile de măsurare a apei de suprafață care măsoară precipitațiile și inundațiile. Mai recent, date relevante pot fi colectate și de la elemente cum ar fi panouri solare, termometre inteligente și aparate de aer condiționat inteligente, legate la internet. În multe sisteme de nowcasting, datele radar sunt utilizate ca sursă de predicție. Chiar și așa, un radar adună date de pe sute de mii de kilometri pătrați, pe mai multe nivele și returnează de zeci de produse diferite care fiecare conține o anumită informație despre vremea curentă. Este greu pentru meteorologii operaționali să analizeze toate aceste date și, de obicei, au un subset de elemente și produse care le consideră cele mai relevante pe care le utilizează pentru generarea de predicții și alerte meteo.

Chiar mai mult, institutele meteorologice dețin un set mare de date meteorologice istorice, cum ar fi măsurători radar, date satelit și observații meteorologice istorice. Aceste date istorice ar putea oferi informații importante pentru șabloanele și manifestările vremii, care ar putea ajuta la crearea unor predicții mai bune. Dar setul de date istorice este prea mare pentru a fi analizat de meteorologi, de aceea sunt necesare sisteme automate. Tehnicile extragere a datelor (data mining) sunt în mod special adecvate pentru astfel de sarcini.

Luând aceste elemente în considerare, am decis să ne apropiem de problema predicției de tip nowcasting în meteorologice din perspectiva învățării automate. Odată ce un model adecvat de învățare

automată este creat și antrenat, acesta poate oferi predicții rapide bazate pe date noi. Un model de învățare automată este antrenat utilizând date existente. În meteorologie există prea multe date istorice pentru fie analizate manual, în timp ce în învățarea automată cu cât există mai multe date de antrenare, cu atât mai bun va fi modelul. Prin urmare, învățarea automată pare să se potrivească pentru prognoza pe pe termen scurt a vremii, deoarece poate profita de cantitatea imensă de date meteorologice istorice. De asemenea, există metode de învățare nesupervizate care pot fi utilizate pentru a analiza datele istorice și pentru a găsi șabloanele și informații importante.

În această lucrare ne propunem crearea modelelor de învățare automată pentru scopuri informative despre vreme. Scopul a fost acela de a crea noi modele care ar putea fi încorporate fie în sistemele existente și deja operaționale sau în sistemele noi, de ultimă oră. Pentru a face acest lucru am ales subsetul datelor meteorologice pentru a fi utilizat pentru dovada conceptului de modele. Am concentrat eforturile noastre asupra datelor radar și am folosit date istorice reale furnizate de Administrația Națională de Meteorologie a României. În timp ce accentul principal a fost acela de a crea modele pentru predicția datelor meteorologice, am creat, de asemenea, modele nesupravegheate pentru analiza datelor, pentru a extrage informații din datele istorice.

Contribuții Originale

Cercetarea noastră a fost focalizată pe două direcții principale: (1) Investigarea modelelor de *învățare nesupervizată* (cum ar fi *rețele cu auto-organizare (Self-Organizing Maps)* și *extragere de reguli de asociere relațională (Relational Association Rules mining)*) pentru analiza datelor meteorologice istorice și extragerea informațiilor; și (2) Dezvoltarea de modele de *învățare supervizată* pentru prognoza pe termen scurt a vremii. Pentru (2) ne-am concentrat pe modelele de *învățare profundă*, cum ar fi *rețele neuronale profunde*, *rețele neuronale recurente* și *rețele neuronale convoluționale*. Și astfel, rezultatele și contribuțiile noastre principale sunt, de asemenea, separate în aceste două direcții, prezentate în capitolele 2 și 3:

(1) **Modele de învățare nesupervizate pentru analiza datelor radar.**

Văzând natura problemei și cantitatea de date istorice disponibile, am considerat că ar fi interesat să analizăm datele pentru a găsi șabloanele sau informații relevante. În acest scop am propus utilizarea *rețelelor cu auto-organizare (Self-Organizing Maps – SOM)* pentru analiza nesupervizată a datelor radar. Aceste experimente au fost făcute urmărind două scopuri principale: elaborarea de metode utile bazate pe SOM-uri pentru analiza datelor radar și extragerea unor informații din datele pe care le-am folosit mai apoi în modelarea predictorului de date radar. Rezultatele noastre pe această linie de cercetare au fost următoarele:

a) Primul nostru experiment a avut scopul de a descoperi modul în care valorile produselor radar evoluează între două scanări ale radarului consecutive. Detaliile și rezultatele muncii noastre în această direcție au fost publicate în [CMtT19]. Metodologia și rezultatele sunt, de asemenea, detaliate în Secțiunea 2.2. Am arătat că, în general, valorile produselor radar se schimbă lent în timp, cu excepția unor momente specifice din cadrul evenimentelor severe. De asemenea, am arătat că datele sunt foarte similare între ele în perioadele în care nu există evenimente meteorologice și că, în timpul evenimentelor semnificative, un produs deosebit de zgomotos (V - Velociate) nu poate fi ignorat, deoarece datele nu sunt la fel de bine descrise fără el.

b) În următorul nostru studiu, am analizat schimbarea valorilor produselor radar la un nivel mult mai scăzut: am căutat șabloane în modul în care valorile se schimbă, pentru un pro-

dus specific, la un pas de timp (moment) specific. Am ales să analizăm un produs extrem de relevant (R02 - Reflectivitatea la al doilea cel mai mic nivel de înălțime), la un moment de timp în mijlocul evenimentului meteorologic sever și la un moment înainte de începerea evenimentului sever. Rezultatele acestei lucrări au fost publicate în [MCt19]. Descriem metodologia și rezultatele în detaliu în secțiunea 2.3 a acestei teze. Am găsit dovezi empirice că valori similare pentru un produs radar la un moment de timp sunt codificate în vecinătăți similare în momentele anterioare, arătând astfel că există o relație semnificativă între valoarea produsului într-un moment și vecinătatea acestuia în momentul precedent, o relație care poate fi utilizată de algoritmi supervizați pentru predicție. Am arătat, de asemenea, că același șablon se este valabil atât în condițiile meteorologice normale, cât și pentru condițiile meteorologice severe și, de asemenea, este valabil și dacă luăm în considerare doar 1 pas de timp sau 5 pași de timp anteriori, arătând că ar putea fi posibil să se creeze predicții doar dintr-un singur pas anterior (făcând antrenarea și predicțiile mai rapide).

c) Întrucât experimentul anterior a fost realizat doar pe un singur produs specific, am vrut să verificăm dacă aceleași șabloane apar și la alte produse radar dintre cele care am considerat să le folosim pentru predicție. Astfel, am extins experimentul pentru a studia, de asemenea, date de la cel mai jos nivel și, de asemenea, alte 2 produse (V - Velocitate și VIL - Lichid integrat vertical (Vertically Integrated Liquid)). Rezultatele acestei extinderi a experimentelor au fost publicate în [Mih20]. În această teză, detaliile și rezultatele sunt prezentate în secțiunea 2.4. Am arătat că aceeași relație și aceleași șabloane apar pentru toate produsele și nivelele de elevație considerate.

(2) **Modele de învățare automată supervizate pentru prognoza pe pe termen scurt a vremii.**

Cea de-a doua direcție de cercetare a noastră este de a dezvolta noi modele supervizate de prognoza pe pe termen scurt a vremii. Scopul a fost de a crea noi modele de învățare automată care pot fi utilizate pentru predicția datelor radar și de a le valida ca dovezi de concept ("proof of concept"). Pentru a valida modelele, am folosit diferite măsuri, dar cele mai relevante sunt *rădăcina pătrată a erorii medii pătratice* (Root Mean Squared Error - RMSE) pentru sarcinile de regresie și *indicele de succes critic* (Critical Success Index - CSI) pentru sarcinile de clasificare. *RMSE* este adesea folosită ca măsură în literatura de predicție a vremii și *CSI* este o măsură specifică meteorologiei, pentru predicțiile care spun dacă va exista un eveniment meteorologic la o locație. De asemenea, am calculat *RMSE* numai pentru valori diferite de zero, deoarece acestea sunt valorile meteorologice relevante, iar valorile zero sunt mult mai multe decât valorile diferite de zero, distorsionând astfel rezultatele. În timpul cercetării noastre am dezvoltat următoarele 3 modele de învățare automată:

a) **NowDeepN.** Primul model pe care l-am creat s-a bazat pe *rețele neuronale profunde*. Ideea a fost de a prezice valoarea unui produs radar într-o locație pe baza valorilor tuturor produselor la pasul de timp anterior într-o vecinătate a acelei locații. Deoarece vrem să prezicem mai multe produse, avem mai multe rețele prevăzute pentru fiecare produs - rezultând un model care conține un ansamblu de 13 rețele. Descrierea și rezultatele modelului *NowDeepN* au fost publicate în [CMt21]. Acestea sunt, de asemenea, descrise în detaliu în Secțiunea 3.1. Pe datele de testare am obținut un *RMSE* de 2,25 $\pm$0,12 cu zerouri și de 5,93 $\pm$0,14 pentru valori diferite de zero. Dacă am considerat valoarea de 5 dbZ ca prag pentru clasificare, am obținut un *CSI* de 0,64. Comparativ cu alte lucrări asociate din domeniu, comparația este favorabilă pentru *NowDeepN* în 5 din 7 cazuri.

b) **RadRAR.** Acest model se bazează pe extragerea *Regulilor de asociere relațională (RAR)*. În timp ce foloseam inițial extragerea de RAR-uri ca instrument nesupervizat de extragere de informații din date, am găsit mai târziu o modalitate de a folosi regulile extrase pentru predicție și am finalizat prin crearea modelului de clasificare binară *RadRAR*. Unul dintre dezavantajele *RAR-urilor* este că acestea sunt mai puțin scalabile, *RadRAR* a fost antrenat și testat pe o regiune geografică mai mică decât celelalte 2 modele și are în vedere doar un produs radar (R01 - Reflectivitatea la unghiul cel mai mic de înălțime). Munca noastră cu privire la acest model este publicată în [CMC19b]. Am descris detaliile acestui model în secțiunea 3.2 din această teză. Folosind un prag de 35 dbZ, care este un prag meteorologic relevant pentru produsul R01, am obținut un *CSI* de 0,56 $\pm$0,02, obținând performanțe mai bune în 8 din 9 comparații cu lucrările asociate din domeniu și cu alți clasificatori.

c) **XNow.** Al treilea model pe care l-am dezvoltat se bazează pe *rețele convoluționale profunde*. De data aceasta, am început cu ideea de a prezice întreaga regiune și toate produsele simultan, din datele din momentul anterior. Modelul este puternic inspirat de arhitecturile *UNet* [RFB15] și *Xception* [Cho17], *XNow* fiind de fapt o versiune modificată a acesteia din urmă pentru a funcționa în mod similar cu prima. Acest model a fost publicat în [SCIM20] și este prezentat în detaliu în secțiunea 3.3 din această teză. Cu modelul *XNow* am obținut un *RMSE* de 1,85 $\pm$0,15 pe date cu zerouri și 2,28$\pm$0,17 pe valori diferite de zero. Acesta este un rezultat foarte bun, fiind mai bun decât *NowDeepN* și marginal mai bun decât cel mai bun model pe care l-am găsit în literatura de specialitate, cu un design și un scop similar.

Structura Tezei

Restul tezei este organizat după cum urmează. În primul capitol este prezentat contextul teoretic și analiza literaturii de specialitate. În Secțiunea 1.1 prezentăm mai întâi tipul de date radar pe care le folosim și modul în care sunt colectate și apoi prezentăm analiza noastră a literaturii. În a doua parte a primului capitol – Secțiunea1.2 – detaliază baza teoretică necesară algoritmilor de învățare automată pe care i-am folosit în cercetarea noastră.

În Capitolul 2 prezentăm experimentele noastre folosind metode de învățare automată nesupervizate – în principal rețele cu auto-organizare – pe date radar. Deoarece toate aceste experimente utilizează același set de date, prezentăm mai întâi în detaliu acest set de date în Secțiunea 2.1. Primul experiment este descris în secțiunea 2.2, unde analizăm schimbarea valorilor produselor radar în timp. Al doilea experiment al nostru folosind rețele cu auto-organizare este detaliat în Secțiunea 2.3. Deoarece cel de-al doilea experiment al nostru a fost realizat doar pe unul dintre produsele radar, la un singur unghi de înălțime, am luat în considerare extinderea experimentului la mai multe produse și unghiuri de înălțime. Rezultatele acestei extensii sunt prezentate în secțiunea 2.4. De asemenea, am introdus noi măsuri de evaluare pentru a interpreta mai bine rezultatele și eficiența SOM-urilor.

Cele trei modele de învățare automată supervizată pe care le-am dezvoltat pentru prognoza pe pe termen scurt a vremii sunt descrise în Capitolul 3. Primul model propus este *NowDeepN*, prezentat în secțiunea 3.1, care se bazează pe rețele neuronale profunde. Al doilea model pe care l-am dezvoltat, bazat pe extragerea regulilor de asociere relațională, este descris în Secțiunea 3.2 a acestei teze. Ultimul model pe care l-am dezvoltat, *XNow*, bazat pe rețele neuronale convoluționale, și mai exact pe arhitectura *Xception*, este prezentat în Secțiunea 3.3.

***

Autorul a beneficiat de sprijin financiar acordat prin finanțare NO Grants 2014-2021, pe baza contractului de proiect nr. 26/2020. Autorul dorește să-i mulțumească lui Eugen Mihuleț de la Administrația Națională de Meteorologie din România, pentru furnizarea seturilor de date meteorologice utilizate în experimente și pentru feedback-ul său util cu privire la rezultatele obținute.

Capitolul 1

Bază teoretică

În acest capitol prezentăm elementele teoretice care vor fi utilizate în teză. Capitolul este împărțit în două părți: prognoza pe pe termen scurt a vremii, acoperind elemente legate de prognoza pe pe termen scurt a vremii și datele radar; și metode de învățare automată, care acoperă elemente privind metodele de învățare automată luate în considerare în cercetarea noastră.

1.1 Prognoza pe pe termen scurt a vremii

Conform unui articol recent comun din țările nordice și baltice [Swe18], sunt așteaptate în viitor schimbări climatice, inclusiv fenomene de ploaie extremă. În consecință, este o nevoie din ce în ce mai mare de avertizare precisă și timpurie a evenimentelor meteorologice severe. Pe măsură ce numărul și intensitatea fenomenelor meteorologice severe crește, prezicerea lor în timp util pentru a evita dezastrele devine extrem de solicitantă pentru meteorologi.

Domeniul din predicția meteo care se ocupă cu analiza și prognozele meteo pentru următoarele de la 0 până la 6 ore se numește *nowcasting – prognoza pe pe termen scurt a vremii* – și joacă un rol din ce în ce mai important în gestionarea crizelor și prevenirea riscurilor. Problema emiterii unor alerte de nowcasting este o sarcină dificilă pentru meteorologi, în principal din cauza setului extrem de mare de date care trebuie analizate într-o perioadă scurtă de timp. Prin urmare, metodele bazate pe învățare automată sunt utile pentru a oferi soluții eficiente pentru prognoza pe pe termen scurt a vremii prin învățarea de șabloane relevante din cantitatea mare de date meteo și îmbunătățind astfel luarea deciziilor în caz de vreme cu impact ridicat. Majoritatea metodelor operaționale și semi-operaționale existente de nowcasting utilizează extrapolarea datelor radar și a algoritmilor bazați în principal pe urmărirea celulelor.

Teza actuală folosește date radar furnizate de un radar meteorologic de tip WSR-98D [NOA18]. Aproximativ la fiecare 6 minute sunt colectate date despre un set complet de aproximativ 30 de produse de bază și produse derivate, adunate pe 7 nivele de înălțime diferite. Produsele de bază sunt *reflectivitatea* particulelor (R), oferind informații despre dimensiunea și tipul particulelor și *velocitatea* particulelor (V), conținând informații despre mișcarea particulelor. Ambele produse sunt disponibile pentru mai multe unghiuri de înălțime ale antenei radar și, pentru fiecare pas de timp, se livrează un set de șapte produse de date, R01-R07 și V01-V07, fiecare dintre ele corespunzând unei anumite înclinări a antenei. Printre produsele derivate, un interes deosebit pentru studiu este VIL (lichid integrat vertical – Vertically Integrated Liquid), o estimare a masei totale a precipitațiilor peste o anumită unitate de suprafață.

1.2 Modele de învățare automată

Învățarea supervizată este un sub-domeniu al învățării automate, care se ocupă de ideea de a aproxima o mapare de la un anumit domeniu de date de intrare la un anumit domeniu de date de ieșire bazat pe exemple de perechi intrare-ieșire. *Învățara nesupervizată* este un sub-domeniu al învățării automate care tratează algoritmi care extrag informații utile din date brute fără a utiliza exemple etichetate (cum se face în învățarea supervizată).

O *rețea cu auto-organizare* (SOM – Self-Organizing Map) [SK99] este un model de învățare nesupervizat, un tip de rețea neuronală artificială din categoria rețelelor de *învățare competitivă*. Un SOM conține două straturi: stratul de intrare și stratul de ieșire. Aceste straturi sunt *dens* conectate. De obicei, un SOM este instruit folosind algoritmul Kohonen [SK99]. Un SOM este un instrument pentru vizualizarea datelor cu dimensiuni ridicate, dar este, de asemenea, și foarte eficient pentru probleme de clusterizare, sarcini de extragere de date sau clasificare [LO92]. Metoda U-Matrix [KK96] este folosită de obicei pentru a vizualiza un SOM antrenat.

textit Regulile de asociere relațională (RAR) [SCC06] sunt o extensie a *Regulilor de asociere* (RA), care sunt instrumente puternice de analiză a datelor și de extragere de informații.

RAR-urile pot descoperi diferite tipuri de relații între atributele datelor. Descoperirea de Reguli de Asociere Relațională ($DRAR$) este algoritmul de tip Apriori folosit pentru extragerea RAR-urilor interesante dintr-un set de date [CBC12].

Metodele de învățare bazate pe *rețele neuronale* oferă o abordare robustă pentru aproximarea funcțiilor țintă cu valoare reală, discretă sau vectorială [Mit97]. Rețelele neuronale sunt potrivite pentru probleme care conțin date zgomotoase și complexe, cum ar fi date de la camere foto/video, microfoane sau senzori. Succesul lor se datorează similitudinii lor cu sistemele biologice eficiente, care sunt capabile să generalizeze și să asocieze date pe care nu au fost antrenate în mod explicit în timpul etapei de antrenare și să coreleze aceste date cu o clasă de care aparțin.

Spre deosebire de rețelele neuronale clasice, *rețelele neuronale profunde* conțin mai multe straturi ascunse și au un număr mare de parametri, ceea ce le face capabile să exprime funcții țintă complicate, i.e mapări complexe între intrările și ieșirile lor [SHK+14]. În prezent, rețelele neuronale profunde sunt modele puternice din literatura de specialitate a învățării automate, aplicate pentru probleme complexe de clasificare și regresie din diferite domenii.

Rețelele neuronale recurente (RNR) [HS13] sunt capabile să exprime procese dinamice și temporale și să modeleze informații secvențiale. Datorită capacității lor de a modela secvențe, RNR au fost folosite cu succes pentru a rezolva numeroase sarcini în care intrarea a fost organizată în pași de timp, inclusiv: recunoașterea vorbirii [Lip15], [GWD14], procesarea imaginilor și a videoclipurilor [WKS16], traducere automată și analiză de sentiment [CvMG+14]. Un RNR conține cel puțin o conexiune de feedback, astfel încât activările pot circula într-o buclă [HS13]. Acest lucru permite rețelelor să efectueze procesare temporală și să învețe secvențe, cum ar fi: recunoașterea/reproducerea secvenței sau asocierea/predicția temporală.

Rețelele neuronale de tip LSTM (Long Shor Term Memory) [HS97] sunt un tip particular de RNR care permit activărilor unitate să păstreze informații importante pe o perioadă mult mai lungă de timp. Pentru a stoca o informație pentru o perioadă mai lungă, trebuie implementat un circuit pentru simularea unei celule de memorie.

Rețelele neuronale convoluționale (RNC) [KSH12] sunt rețele neuronale artificiale care primesc ca intrare imagini cu mai multe canale. Deoarece acestea sunt rețele neuronale la bază, conceptele rețelelor neuronale sunt valabile și pentru RNC: primesc o intrăre, o procesează prin funcția de propagare, apoi transmit rezultatul la o funcție de activare, și în cele din urmă produce o ieșire.

Ideea rețelelor de tip ConvLSTM este de a introduce convoluții în interiorul celulei LSTM.

LSTM-urile nu folosesc date spațiale, ci doar date temporale, ceea ce reprezintă un mare dezavantaj pentru sarcinile care ar putea profita de datele spațiale. Combinația de LSTM și RNC a fost menționată în 2014 de Donahue și colab. în [DHG+14].

Capitolul 2

Noi modele de învățare nesupervizată pentru analiza datelor meteorologice

Predicția vremii și, în special, a vremii severe, este o provocare importantă atât pentru cercetătorii din domeniul meteorologiei, cât și pentru cei din domeniul învățării automate. Complexitatea și dificultatea problemei se datorează în principal caracterului haotic al atmosferei, cât și setului implicit de informații meteorologice (radar, satelit sau observații meteorologice terestre) care trebuie analizate de meteorologi. Astfel, înțelegerea relațiilor dintre diferiți parametri meteorologici extrași din observațiile făcute de radar radar poate fi utilă pentru a oferi o înțelegere mai bună asupra evoluției vremii severe și ar ajuta la identificarea situațiilor în care poate apărea vreme severă.

În secțiunile următoare, *rețelele cu auto-organizare* sunt explorate ca un model de clasificare nesupervizat pentru detectarea șabloanelor din datele radar care sunt relevante în prezicerea schimbărilor meteo pe termen scurt.

În studiile prezentate în acest capitol au fost utilizate date reale furnizate de Administrația Națională de Meteorologie (ANM). În Secțiunea 2.1 acest set de date este prezentat în detaliu.

Toate elementele prezentate în acest capitol au fost, de asemenea, publicate în trei lucrări originale [CMtT19, MCt19, Mih20]. În cele ce urmează subliniem principalele contribuții originale prezentate în capitol:

- Primul experiment este prezentat în Secțiunea 2.2. Cu scopul principal de a analiza modul în care valorile produselor radar meteorologice evoluează între scanări radar consecutive, arătăm empiric că, în general, există o schimbare lentă a valorilor în timp, cu excepția situațiilor în care apar anumite fenomene severe. Studiul realizat în Secțiunea 2.2 este o lucrare originală publicată în [CMtT19] și are scopul de a oferi o perspectivă mai bună cu privire la modul în care valorile produselor radar meteorologice evoluează în timp atât în condiții meteorologice calme, cât și în condiții meteorologice severe; având obiectivul mai larg al utilizării acestor rezultate pentru prognoza pe pe termen scurt a vremii.

- Secțiunea 2.3 introduce un model alternativ de date radar și vizează obținerea unei dovezi empirice că: (1) există unele șabloane în modul în care valorile produselor radar trec de la un moment de timp la altul, atât în condiții meteorologice normale, cât și în condiții meteorologice severe; și (2) că valori similare pentru un produs la un moment de timp sunt codificate în vecinătăți similare în momente de timp anterioare. Abordarea din această secțiune a fost publicată în lucrarea originală [MCt19].

- Experimentul anterior a fost, la început, testat pe un singur produs radar. Am extins apoi domeniul de aplicare al studiului pentru alte produse radar cu scopul de a susține ideea că re-

zultatele sunt consistente pentru diferite produse de date radar. Această extensie este prezentată în Secțiunea 2.4, iar rezultatele au fost publicate în [Mih20].

Prezentarea din acest capitol se bazează pe lucrările originale [CMtT19, MCt19, Mih20].

2.1 Setul de date radar

Pentru experimentele noastre folosim date radar reale furnizate de ANM, administrația meteorologică din România.

Datele au fost furnizate de un radar meteorologic de tip WSR-98D [NOA18] situat în Bobohalma, România. Aproximativ la fiecare 6 minute sunt colectate date despre un set complet de aproximativ 30 de produse de bază și produse derivate, adunate pe 7 nivele de înălțime diferite. Produsele de bază sunt *reflectivitatea* particulelor (R), oferind informații despre dimensiunea și tipul particulelor și *velocitatea* particulelor (V), conținând informații despre mișcarea particulelor. Ambele produse sunt disponibile pentru mai multe unghiuri de înălțime ale antenei radar și, pentru fiecare pas de timp, se livrează un set de șapte produse de date, R01-R07 și V01-V07, fiecare dintre ele corespunzând unei anumite înclinări a antenei. Printre produsele derivate, un interes deosebit pentru studiu este VIL (lichid integrat vertical – Vertically Integrated Liquid), o estimare a masei totale a precipitațiilor peste o anumită unitate de suprafață. Grila de date furnizată de radar pentru zona geografică selectată la un moment dat se potrivește cu o matrice. Radarul oferă o matrice de date pentru fiecare dintre produsele meteorologice și fiecare matrice are 624 de rânduri și 800 de coloane (adică $m = 800$ și $n = 624$).

Ziua utilizată ca studiu de caz este 5 iunie 2017, o zi cu instabilitate atmosferică moderată în regiune, manifestată prin furtuni însoțite de ploi abundente și grindină de dimensiuni medii. În ceea ce privește aceste fenomene, Administrația Națională de Meteorologie a emis cinci alerte meteorologice severe, cod galben. În zona geografică aleasă, au existat două episoade distincte cu evenimente meteorologice intense în 5 iunie 2017: primul a avut loc aproximativ între 09:00 și 11:00 UTC, iar al doilea aproximativ între 12:00 și 17:00 UTC, cele mai severe evenimente având loc între orele 14:00 și 15:00 UTC.

Datele radar utilizate în studiul nostru de caz au fost înregistrate între 00:04:04 UTC și 23:54:02 UTC. Avem în total 231 de momente de timp (adică $k = 231$), cu momentul 1 de timp corespunzătoar cu 00:04:04 UTC. Cele mai interesante momente de timp sunt cele în care există date despre evenimentele meteorologice menționate mai sus: momentele de timp de la 88 la 106 conțin datele pentru evenimentul meteorologic de la 09:00 la 11:00 și momentele de timp de la 117 la 165 conține datele pentru evenimentul meteorologic de la 12:00 la 17:00. Datele pentru valorile maxime, având loc aproximativ între 14:00 și 15:00, sunt conținute în momentele de timp de la 137 la 145.

Datele colectate de radar conțin o valoare specială care reprezintă „Fără date”. Această valoare este de obicei reprezentată de -999, dar am decis să o înlocuim cu 0, deoarece în majoritatea cazurilor această valoare se referă la particulele de aer cu 0 reflectivitate (adică fără picături de apă semnificative). „Fără date” poate reprezenta, de asemenea, volume de aer care nu au returnat niciun semnal, de exemplu dacă un sector cu reflectivitate ridicată se află între radar și locația respectivă. În acest caz, înlocuirea acestuia cu 0 este, de asemenea, corectă, deoarece întreaga regiune este obturată și datele nu sunt relevante pentru procesul de învățare. Datele radar sunt predispuse la diferite tipuri de erori, meteorologice și tehnice, care implicit se găsesc în matricea de date de ieșire. Erorile meteorologice (de exemplu, subestimarea reflectivității unei particule) sunt dificil de identificat și eliminat, dar unele erori care apar în timpul conversiei datelor au fost identificate și corectate. De exemplu, produsul V trebuie să conțină doar valori între -33 și 33, dar în datele noastre am găsit valori de -100. Pentru a evita introducerea de în experimentele noastre a erorilor pe care le reprezintă aceste valori, am decis

să le omitem în procesul de învățare nesupervizată. Mai exact, în timpul antrenamentului utilizând algoritmul Kohonen, valorile eronate de -100 au fost omise în timpul calculării distanței euclidiene dintre instanțele de intrare și neuronii de pe hartă.

2.2 Analiza schimbării datelor radar în timp

Rețelele cu auto-organizare sunt explorate în această secțiune ca un model de clasificare nesupervizată pentru detectarea șabloanelor din datele radar care sunt relevante în predicția pe termen scurt a schimbărilor vremii. Abordarea introdusă în această secțiune este o lucrare originală publicată în [CMtT19]. Cu scopul principal de a analiza modul în care valorile produselor radar evoluează între scanări ale radarului consecutive, arătăm empiric că, în general, aceste valori se modifică încet în timp, cu excepția situațiilor în care apar anumite fenomene severe. Studiul realizat, prezentat în această secțiune, are scopul de a oferi o perspectivă mai bună cu privire la modul în care valorile produselor radar evoluează în timp, atât în condiții de vreme calmă, cât și în condiții de vreme severă, cu scopul mai larg de a utiliza aceste descoperiri pentru prognoza pe pe termen scurt a vremii.

Evaluăm utilitatea SOM-urilor pentru a descoperi nesupervizat structura de bază a datelor radar, pentru a analiza modul în care valorile pentru mai multe produse radar meteorologice evoluează între scanări radar consecutive și pentru a studia relevanța produselor radar în prognoza pe pe termen scurt a vremii. Prin mai multe experimente efectuate pe date radar reale, furnizate de ANM, ne propunem să obținem o dovadă empirică că valorile produselor meteorologice ale radarului se modifică relativ încet în timp în condiții meteorologice normale, cu excepția situațiilor în care apar anumite fenomene severe . În plus, ne așteptăm ca SOM-urile să poată distinge dintre condițiile meteorologice severe și cele normale pe baza datelor radar.

În cele ce urmează propunem un model de date care va fi utilizat în continuare în experimentele noastre. Ideea este de a atribui, la fiecare moment de timp, o reprezentare vectorială fiecărei grile de date 3D furnizate de radar. În acest model, pentru o zi d, un moment de timp t_i^d $(1 \leq i \leq k)$ și un set $Prod$ de produse meteorologice, un paralelipiped de date $P_{t_i{}^d}(m, n, Prod) = (p_{xyz})_{\substack{x=\overline{1,m} \\ y=\overline{1,n} \\ z=\overline{1,|Prod|}}}$ este construit. În acest paralelipiped, axele OX și OY reprezintă rândurile și coloanele din grila de date radar, iar axa OZ reprezintă produsele meteorologice. Pentru obținerea reprezentării vectoriale, acest paralelipiped $P_{t_i{}^d}(m, n, Prod)$ este liniarizat.

Două seturi de date, $D1$ și $D2$, sunt construite pentru a reprezenta datele radar colectate în momentele de timp $t_1, t_2,\ dots, t_k$ utilizând modelul de date introdus anterior. Diferența dintre $D1$ și $D2$ este dată de setul de produse meteorologice utilizate pentru reprezentarea instanțelor. În $D1$ se utilizează întregul set de produse meteorologice furnizate de radar (adică 24), în timp ce $D2$ folosește doar 13 produse:*reflectivitatea* (R) particulelor pe șase nivele, *velocitatea* (V) particulelor, pe șase nivele, și cantitatea estimată de apă conținută de o coloană de aer de un metru pătrat (VIL).

Pentru detectarea structurii de la baza seturilor de date $D1$ și $D2$, se aplică modelul SOM pentru a obține în mode nesupervizat o reprezentare bidimensională a celor două seturi de date.

Ca și etapă preliminară înainte de aplicarea modelelor SOM nesupervizate, a fost efectuată o analiză statistică a setului de date cu scopul de a analiza variația produselor meteorologice pe fiecare moment de timp.

Pentru SOM-urile utilizate în experimente am folosit propria noastră implementare, fără a utiliza alte biblioteci. Pentru construirea SOM-ului, am folosit o topologie tip tor [KTO+07].

Parametri folosiți pentru SOM sunt următorii: o configurație de 30x30 de neuroni pe hartă, 20000 epoci de antrenament și o rată de învățare de 0.1.

În implementarea noastră, pe U-matrice valorile mai mici sunt marcate ca locuri mai întunecate, în timp ce valorile mai mari sunt marcate ca regiuni mai albe. În consecință, regiunile mai întunecate codifică instanțe de date similare, în timp ce regiunile mai albe reprezintă limite de separare între grupurile de date.

Rezultatele sugerează că ar putea exista unele modificări vizibile în produsele meteorologice cu aproape 2 ore înainte de începerea evenimentului, care ar putea ajuta la predicția începerii fenomenelor.

De asemenea, am concluzionat că valorile R, V și VIL pot fi utile pentru prezicerea evenimentelor meteorologice și că produse meteorologice adiționale (altele decât R, V și VIL) nu aduc informații semnificative despre fenomene.

Pentru evaluarea relevanței produsuluiV în analiza nesupervizată a evenimentelor meteorologice, am efectuat primul experiment folosind doar produse R și VIL, fără a lua în considerare V. Analizând rezultatele am ajuns la concluzia că și produsul V este relevant în analiza evenimentelor meteorologice severe, iar măsurile R și VIL trebuie utilizate împreună cu V pentru a crește performanța procesului.

Rezultatele arată dovezi că valorile produselor radar discriminează în mod clar între vremea calmă și evenimentele severe. SOM-ul este, de asemenea, capabil să detecteze nesupervizat aceste șabloane folosind doar produsele R, V și VIL. Acest lucru sugerează fezabilitatea învățării de a prezice (folosind produse R, V și VIL) un întreg paralelipiped de date la un anumit moment, pe baza paralelipipedelor de date din momentele anterioare.

Această secțiune a prezentat un studiu privind aplicarea SOM-urilor ca metodă de clasificare nesupervizată pentru analiza datelor radar meteorologice și investigarea relevanței mai multor produse meteorologice în detectarea fenomenelor meteorologice severe. În condiții meteorologice normale, valorile produselor meteorologice se schimbă ușor în timp, cu excepția situațiilor în care apar anumite fenomene severe. Astfel, evenimentele meteorologice reflectate în modificările survenite în valorile mai multor produse meteorologice sunt într-adevăr detectate de algoritmi de învățare nesupervizați.

2.3 Analiza șabloanelor din tranzițiile datelor dintre scanări consecutive ale radarului

Abordarea introdusă în această secțiune este o lucrare originală publicată în [MCt19].

Scopul principal al abordării introduse în această secțiune este de a înțelege mai bine relațiile dintre produsele meteorologice rezultate dintr-o observație radar și unele observații de date radar din momente de timp anterioare, atât în condiții meteorologice severe, cât și normale.

În această secțiune investigăm capacitatea modelelor SOM de a învăța nesupervizat șabloane meteorologice relevante, în special în situațiile în care au avut loc evenimente meteorologice severe. Ne concentrăm în mod special pe șabloane ca urmare a tranziționării de la un moment de timp la altul. Ca o dovadă a conceptului, SOM-urile au fost folosite în studiul actual ca instrument de învățare nesupervizată pentru analiza datelor radar prelevate la nivel național și utilizate pentru prognoza pe pe termen scurt a vremii. Prin experimente, oferim o dovadă empirică că (1) există unele șabloane în modul în care valorile produselor radar trec de la un moment de timp la altul, în condiții meteorologice normale cât și severe, și că (2) valori similare pentru un produs la un moment de timp dat sunt codificate în vecinătăți similare în momentele de timp anterioare.

Pentru modelarea computațională a datelor radar, propunem un model de date *la nivel de celulă*. În acest model, ne propunem să atribuim unei anumite celule (x, y) din grilă, la fiecare moment de timp $timp$ o reprezentare vectorială. Această reprezentare conține valorile produselor de date dintr-o zonă vecină (subgrilă) cu o anumită lungime care înconjoară punctul (x, y), pentru o secvență

temporală de lungime l înainte de $timp$: $timp$ - l, $timp$ - l + 1, . . . , $timp$ - 1.

Experimentele au scopul de a analiza măsura în care SOM-urile sunt capabile să descopere nesupervizat fenomenele meteorologice în datele radar. Scopul experimentelor este de a testa dacă valori similare pentru un produs radar la un moment dat sunt codificate în vecinătăți similare în momentele anterioare. Pentru un anumit moment de timp t, sunt construite două seturi de date D și D'. Diferența dintre D și D' este dată de lungimea l considerată pentru secvența temporală: în D folosim o lungime l de 1, în timp ce D' consideră o valoare pentru l mai mare de 1.

Ca etapă preliminară înainte de aplicarea modelului SOM nesupervizat, a fost efectuată o analiză statistică asupra setului de date, cu scopul de a analiza variația produselor meteorologice în fiecare moment de timp.

Pentru SOM-urile [SK99] utilizate în experimente am folosit propria noastră implementare, fără a utiliza alte biblioteci. Pentru construirea SOM-ului, am folosit o topologie tip tor [KTO+07]. În implementarea noastră, valorile mai mici din U-matrice sunt marcate ca și locuri mai închise la culoare, în timp ce valorile mai mari sunt marcate ca regiuni mai deschise. În consecință, regiunile mai închise codifică instanțe de date similare, în timp ce regiunile mai albe reprezintă limite de separare între clustere. Prin acest experiment, ne-am aștepta ca SOM-ul să detecteze nesupervizat o relație între valoarea unui anumit produs pentru o celulă c la un anumit moment de timp t și reprezentarea sa vectorială utilizând model de date *la nivel de celulă* propus la momentele de timp care preced t.

Analiza rezultatelor a condus la concluzia că valori mari pentru $R02$ pot fi prezise din momentele de timp anterioare, indiferent de lungimea secvenșei temporale și indiferent dacă există sau nu evenimente meteorologice severe. În plus, este posibil de prezis o estimare suficient de bună a valorii reale a $R02$. Acest rezultat implică faptul că există șabloane ce pot fi învățate din date, iar învățarea supervizată pentru predicția de date $R02$ este fezabilă în scopul prognozei pe pe termen scurt a vremii.

Ca o concluzie a studiului nostru, SOM-urile sunt capabile să descopere nesupervizat șabloane ascunse în datele radar, care sunt relevante dintr-o perspectivă meteorologică. Rezultatele studiului nostru sugerează rezultate promițătoare în aplicarea modelelor predictive de învățare supervizată pentru prognoza pe pe termen scurt a vremii folosind date radar.

2.4 Extinderea analizei pentru mai multe produse radar

În secțiunea anterioară (Secțiunea 2.3) ne-am concentrat experimentele pe un singur produs radar, R02.

În această lucrare, publicată în [Mih20], extindem și analizăm în continuare capacitatea SOM-urilor de a codifica și extrage din date radar șabloane relevante legate de la modul în care produsele radar se schimbă de la un moment de timp la altul, analizând alte patru produse radar și arătând empiric că rezultatele noastre anterioare pot fi generalizate pentru cele mai utilizate produse radar în prognoza pe pe termen scurt a vremii.

Înainte de a construi modelul SOM, se aplică mai întâi un pas de *curățare* a datelor radar. Scopul acestui pas de preprocesare este de a corecta valorile eronate furnizate de radar. Valorile eronate reprezintă valori care sunt în afara limitelor pentru un produs (de exemplu, o valoare de 75 pentru R01, care în mod normal ar trebui să fie între 0 și 65). Corectăm aceste valori folosind un algoritm de estimare care estimează valoarea corectă din valorile punctelor din vecinătatea 13x13 din jurul valorilor eronate.

Principalul model de învățare automată utilizat este SOM-ul, cu o hartă 2D. Fiecare instanță a datelor de intrare este un vector la *nivel de celulă*, fiind același model de date ca cel utilizat în secțiunea anterioară. În această lucrare extindem acel experiment pentru a investiga dacă ipoteza că valori similare la un moment dat sunt codificate în vecinătăți similare în momentele anterioare încă

se păstrează pentru alte produse de date radar ($R01$, $V01$, $V02$ și VIL). În experimentele noastre se utilizează două lungimi diferite ale secvenței temporale, secvența de momente de timp anterioare folosite: numai *un* moment de timp anterior și *cinci* momente de timp anterioare. Prin urmare, avem opt rezultate experimentale - pentru fiecare dintre cele patru produse radar ($R01$, $V01$, $V02$ și VIL) avem două rezultate, unul cu un moment de timp anterior și unul cu cinci momente de timp anterioare.

Pentru evaluarea calității mapării SOM-ului, introducem o măsură de evaluare ASE (*eroarea medie de similaritate*) care măsoară cât de asemănătoare sunt valorile unui anumit produs radar care sunt mapate pe regiuni similare din SOM. Introducem această eroare pentru a măsura cât de diferită este maparea reală pe un neuron de maparea ideală.

Măsura ASE va avea valori cuprinse între 0 și 1, unde 0 înseamnă că toți neuronii interesanți au etichete de valoare egală (ceea ce este ideal) și 1 înseamnă că toți neuronii interesanți au etichete de la ambele extreme în cantitate egală. Prin urmare, valori mai mici pentru ASE indică o mapare mai bună, din punct de vedere meteorologic.

Pentru a obține o perspectivă mai bună asupra structurii datelor radar, am decis să înăsprim constrângerile impuse pentru califcarea de *neuron interesant*. Am decis să facem acest lucru pentru că am observat că au existat mulți neuroni care aveau multe etichete 0 și una sau foarte puține etichete diferite de zero, dar foarte aproape de 0. Prin urmare, avem o *eroarea medie de similaritate* secundară (ASE') care ia în considerare numai neuronii care conțin doar valori diferite de zero.

Am folosit în experimentele noastre propria noastră implementare pentru modelul SOM, care a fost construită folosind o rețea 2D având o topologie tip tor [KTO+07]. Metoda U-matrice este utilizată pentru a vizualiza maparea rezultată, unde valorile mai mici din U-matricea sunt descrise ca regiuni mai întunecate, în timp ce regiunile mai albe prezintă valori mai mari.

Analizând U-matricile rezultate din cele 8 experimente efectuate, am observat că acestea sunt în concordanță cu rezultatele obținute în lucrarea noastră anterioară [MCt19]. Aceasta înseamnă că combinații similare de valori ale produselor de date radar la momente de timp anterioare sunt corelate cu valori similare ale valorilor produselor la momentul de timp curent, pentru toate produsele studiate, ceea ce duce la concluzia că produsele radar studiate pot fi prezise, momentele de timp viitoare fiind prezise pe baza valorile produselor radar la momentele de timp anterioare.

Pentru fiecare dintre aceste mapări SOM rezultate, am calculat măsurile introduse anterior. ASE este foarte scăzut pentru toate hărțile rezultate. Toate valorile sunt sub 0,05, cu excepția R01, care este sub 0,1. O eroare mai mică de 0,05 înseamnă că etichetele mapate pentru un singur neuron au, în medie, diferențe nu mai mari de 5% din diferența maximă posibilă pentru acea etichetă. Aceasta înseamnă că etichetele mapate la un neuron sunt foarte asemănătoare, ceea ce este de dorit.

Măsura ASE' este foarte asemănătoare cu măsura ASE, singura diferență fiind neuronii pe care sunt măsurate erorile. ASE' se măsoară folosind mult mai puțini neuroni, o treime până la o șesime din numărul de neuroni folosiți de ASE. Cu toate acestea, ASE' nu este mult mai mare decât ASE. Per total, valorile măsurilor ASE și ASE' sunt destul de promițătoare, susținând interpretarea hărților că etichetele similare sunt mapate în regiuni similare.

Folosirea doar a unui moment de timp anterior sau a mai multor momente de timp anterioare pentru antrenament nu pare să aibă un impact semnificativ asupra rezultatului. Folosind 5 momente de timp anterioare, numărul de neuroni utilizați (atât în N cât și în N') a fost mai mic pentru toate experimentele, dar măsurile nu au fost afectate, deoarece sunt foarte asemănătoare.

Capitolul 3

Contribuții în dezvoltarea de modele de învățare profundă pentru prognoza pe termen scurt a vremii

Al doilea scop al cercetării noastre a fost de a crea modele noi de *predicție* pentru prognoza pe pe termen scurt a vremii. Mai exact, am creat modele de învățare automată *supervizată* care să prezică ecoul radar pentru de la moment de timp pe baza momentelor de timp anterioare, și, de asemenea, am validat aceste modelele. Acest capitol prezintă aceste modele de învățare automată *supervizată* și experimentele pe care le-am efectuat.

Toate elementele prezentate în acest capitol au fost, de asemenea, publicate în trei lucrări originale: [CMt21, CMC19b, SCIM20]. Contribuțiile noastre originale prezentate în acest capitol sunt următoarele:

- În Secțiunea 3.1 e prezentat primul nostru model, *NowDeepN*, publicat în [CMt21]. Acest model a fost bazat pe *rețele neuronale profunde*. Ideea a fost de a prezice valoarea unui produs radar, la o anumită locație, pe baza valorilor tuturor produselor la pasul de timp anterior într-o vecinătate a acelei locații. Deoarece prezicem mai multe produse, avem mai multe rețele neuronale pentru fiecare produs Pe datele de test am obținut un *RMSE* de 2,25 $\pm$0,12 cu zerouri și de 5,93$\pm$0,14 pentru valori diferite de zero. Dacă am considera valoarea de 5 dbZ ca prag pentru clasificare, am obținut un *CSI* de 0,64. Comparativ cu lucrările asociate din literatură, comparația este favorabilă pentru *NowDeepN* în 5 din 7 cazuri.

- Secțiunea 3.2 descrie următorul nostru model, *RadRAR*, bazat pe extragerea *regulilor de asociere relațională (RAR)*. Acest model și experimentele conexe au fost publicate în lucrarea noastră [CMC19b]. Deoarece unul dintre dezavantajele *RAR*-urilor este că acestea sunt mai puțin scalabile, *RadRAR* a fost antrenat și testat pe o regiune geografică mai mică decât celelalte 2 modele și are în vedere doar un produs radar (R01 - Reflectivitatea la cel mai mic nivel de înălțime). Folosind un prag de 35 dbZ, am obținut un *CSI* de 0,56$\pm$2, obținând rezultate mai bune în 8 din 9 comparații cu lucrările asociate din literatură și cu alți clasificatori.

- Ultimul model pe care l-am dezvoltat este *XNow*, prezentat în Secțiunea 3.3. Acest model se bazează pe *rețele convoluționale profunde* și a fost publicat în [SCIM20]. Am început cu ideea de a prezice întreaga regiune și toate produsele simultan, din datele din momentul de timp anterior. Modelul este puternic inspirat de arhitecturile *UNet* [RFB15] și *Xception* [Cho17]. Cu modelul *XNow* am obținut un *RMSE* de 1,85 $\pm$0,15 pe date cu zerouri și 2,28$\pm$0,17 pe valori

diferite de zero. Acesta este un rezultat foarte bun, fiind mai bun decât *NowDeepN* și marginal mai bun decât cel mai bun model pe care l-am găsit în literatura de specialitate, care să aibă un un design asemănător și un scop similar.

Prezentarea din acest capitol se bazează pe lucrările originale [CMt21, CMC19b, SCIM20].

3.1 NowDeepN: o abordare pentru prognoza pe pe termen scurt a vremii folosind rețele neuronale profunde

Cu scopul de a ajuta meteorologii să analizeze datele radar pentru emiterea alertelor tip nowcasting, am introdus în lucrarea noastră originală [CMt21] un model de învățare supervizată $NowDeepN$, bazat pe un ansamblu de *rețele neuronale profunde* pentru prezicerea valorilor produselor meteorologice radar care pot fi utilizate pentru prognoza pe pe termen scurt a vremii. Prezentarea modelului din această teză se bazează pe lucrarea noastră publicată [CMt21].

Ca și dovadă de concept, $NowDeepN$ este un model propus să învețe să aproximeze o funcție care face legătura între valorile anterioare ale produselor radar și valorile viitoare ale acestora. Experimentele au fost efectuate pe date radar reale furnizate de Administrația Națională de Meteorologie a României (ANM), colectate din regiunea Transilvaniei centrale.

Pentru *NowDeepN* folosim același model de date introdus în Secțiunea 2.3.

Datele radar sunt predispuse la diferite tipuri de erori, meteorologice și tehnice, care implicit se găsesc în matricea de date de ieșire. Pentru reducerea zgomotului pe care îl reprezintă aceste valori nevalide, este propus un pas de *curățare a datelor*. Ideea care stă la baza etapei de curățare este de a înlocui valorile nevalide dintr-un anumit punct (i, j) cu media ponderată a valorilor dintr-o vecinătate cu lungimea de 13 care înconjoară punctul. Ponderea asociată unui anumit vecin al punctului este invers proporțională cu distanța euclidiană dintre acel vecin și punct, astfel încât valorile celor mai apropiați vecini au o importanță mai mare în estimarea valorii punctului.

Problema de regresie pe care ne-am focalizate este următoarea: să prezicem o succesiune de valori dintr-un un set $Prod$ de produse radar la un moment dat t pentru o anumită locație (i, j) pe hartă , luând în considerare valorile pentru locațiile învecinate punctului (i, j) în momentul t-1. $NowDeepN$ folosește un ansamblu de rețele neuronale profunde pentru a învăța să prezică valorile produselor radar din setul $Prod$ pe baza valorilor lor istorice. Ansamblul este format din np rețele neuronale ($np = |Prod|$), câte o rețea pentru fiecare produs radar.

Una dintre dificultățile legate de problema de regresie formulată anterior este că seturile de date de antrenament sunt extrem de *dezechilibrate*. Mai precis, există multe instanțe de antrenament etichetate cu zero (adică $y_k = 0$) care corespund punctelor de pe hartă fără evenimente meteorologice specifice, și un număr mult mai mic de instanțe cu o etichetă diferită de zero (adică corespunzătoare unui fenomen meteorologic). Natura dezechilibrată a datelor poate duce la un regresor care este părtinitor în favoarea a prezicerii de valori zero, deoarece majoritatea exemplelor de instruire utilizate pentru construirea regresorului au fost etichetate zero.

Pentru evaluarea performanței $NowDeepN$, se aplică o metodologie de testare de tip *validare încrucișată* pe fiecare set de date. Seturile de date sunt împărțite aleatoriu în 5 seturi. Ulterior, 4 seturi vor fi utilizate pentru antrenament și restul de seturi pentru testare, iar acest lucru se repetă pentru fiecare strat (i.e. de 5 ori).

Pentru fiecare împărțire date test-date antrenament, sunt utilizate și calculate două măsuri de evaluare pentru fiecare împărțire de testare a antrenamentului: *eroarea RMSE* (root mean square error – rădăcina pătrată a erorii medii pătratice) și *eroarea NRMSE* (RMSE normalizat) [HK06]. RMSE calculează rădăcina pătrată a mediei erorilor pătrate obținute pentru instanțele de testare. NRMSE

reprezintă RMSE normalizat, obținut prin împărțirea valorii RMSE la intervalul de ieșire și este de obicei exprimat ca procent. Pentru o evaluare mai precisă a rezultatelor, valorile pentru măsurile de evaluare (RMSE și NRMSE) sunt calculate și doar pentru instanțele etichetate cu valor non-zero (RMSE$_{non-zero}$, NRMSE$_{non-zero}$).

Setul de date utilizat în experimentele *NowDeepN* este același cu cel prezentat în 2.1.

Pentru a estima impactul etapei de curățare a datelor, am analizat setul de date înainte și după curățare. Observațiile pe care le-am făcut din această analiză ne conduc la ipoteza că etapa de curățare ar avea impact asupra performanței generale a $NowdDeepN$ și acest lucru ar trebui să fie vizibil cel puțin la nivelurile mai mici de înălțime pentru V.

Pentru rețelele neuronale profunde utilizate în experimentele noastre, a fost utilizată platforma Keras Deep Learning API [Ker18] folosind în spate platforma de rețele neuronale Tensorflow. Codul este disponibil public la [CMt21]. Dat fiind faptul că datele noastre aveau o dimensionalitate relativ ridicată, am avut nevoie de o rețea neuronală relativ complexă. Aceste rețele au fost antrenate pentru 30 de epoci folosind 1024 de instanțe într-unlot de antrenare.

Ne propunem să analizăm cât de corelate sunt rezultatele noastre computaționale cu dovezile meteorologice. Pentru a permite o interpretare mai ușoară a rezultatelor dintr-o perspectivă meteorologică, am calculat *media erorilor absolute* pentru toate instanțele (MAE), cât și numai pentru instanțele marcate non-zero (MAE $_{non-zero}$). Am obținut un NRMSE mediu mai mic de 4% pentru produsele R, ceea ce ar implica o asemănare strânsă între datele prezise și datele reale. Din punct de vedere meteorologic, MAE atât pentru toate instanțele cât și pentru cele diferite de zero este unul satisfăcător, ceea ce înseamnă că valoarea prezisă este la același nivel sau la un nivel apropiat pe scara valorilor produsului.

Pentru a valida empiric ipoteza că etapa de curățare îmbunătățește performanța predictivă a $NowDeepN$, am evaluat modelul instruit pe setul de date necurățat, folosind aceeași metodologie.

Comparând rezultatele, am observat o îmbunătățire a performanței predictive a $NowDeepN$ pentru datele curățate. Pentru a determina semnificația atributelor, comparăm rezultatele $NowDeepN$ folosind setul original de atribute cu rezultatele obținute prin aplicarea $NowDeepN$ pe setul de date după aplicarea a unei etape de extragere a atributelor. S-au folosit două metode de extragere a atributelor pe setul original de atribute, pentru reducerea dimensionalității datelor de intrare: un AE și algoritmul PCA. Comparând rezultatele cu cele obținute fără a aplica o etapă de extragere a atributelor, am observat o îmbunătățire a performanței predictive a $NowDeepN$ pe setul de fără o etapă prealabilă de extragere a atributelor. Relevanța atributelor este validată de faptul că o tehnică de reducere a dimensionalității (AE / PCA) aplicată înainte de predicție cu $NowDeepN$ nu îmbunătățește performanța de învățare.

Am început comparația între $NowDeepN$ și lucrările asociate din literatură prin compararea modelului nostru cu un model de bază simplu: *regresia liniară* (LR). Pentru o comparație exactă, modelul de date utilizat pentru $NowDeepN$ a fost folosit și pentru modelul LR. Prin aplicarea LR pe setul de date s-a obținut un RMSE global pentru valorile diferite de zero (RMSE$_{non-zero}$) de $6,094$.

Am găsit patru abordări având un scop similar cu lucrarea noastră: prezicerea valorilor viitoare ale produselor radar pe baza valorilor lor istorice. Abordările din literatura de specialitaste care sunt cele mai similare cu ale noastre sunt cele propuse de Yan Ji [Ji17], Han et al. [HSZ+17, HSZ19] și Yan și colab. [YJM+20].

Rezultatele arată că, în general, în **71%** din cazuri (5 din 7 comparații), comparația este favorabilă pentru $NowDeepN$. Propunerea noastră este depășită doar de munca lui Yan Ji [Ji17], care a raportat un HR mai bun și un RMSE maxim puțin mai bun decât al nostru.

Tran și Song [TS19] au abordat problema prognozei pe pe termen scurt a vremii din perspectiva procesării de imagini, aplicând anumite praguri asupra valorilor de reflectivitate (20/20/40 dBZ). Re-

zultatele comparative evidențiază faptul că $NowDeepN$ a obținut rezultate mai bune decât modelul propus de Tran și Song [TS19] în **77,7%** din cazuri (7 din 9 comparații). Subliniem performanța mai bună a $NowDeepN$ la valori mai mari a pragului, care indică capacitatea modelului nostru de a detecta precipitații moderate și abundente și grindină medie și mare.

Am introdus în această secțiune un model de regresie bazat pe învățare swupervizată: $NowDeepN$; care utilizează un ansamblu de *rețele neuronale artificiale profunde* pentru prezicerea valorilor produselor meteorologice într-un anumit moment de timp pe baza valorilor lor istorice. $NowDeepN$ se intenționează a fi o dovadă de concept pentru fezabilitatea învățării aproximării unei funcții care face legătura între valorile anterioare ale produselor radar și valorile viitoare ale acestora.

3.2 $RadRAR$: O abordare bazată pe reguli de asociere relațională pentru predicția pe pe termen scurt a datelor radar

Regulile de asociere relațională (RAR) [SCC06] extind clasicele *reguli de asociere* prin integrarea relațiilor dintre valorile atributelor care caracterizează un set de date. În lucrarea noastră originală [CMC19b] investigăm, ca dovadă de concept, relevanța aplicării extraerii RAR-urilor din date cu scopul de a face distincția între condițiile meteorologice severe și cele normale, pentru a folosi aceste predicții în prognoza pe pe termen scurt a vremii. În plus, ne propunem să subliniem relevanța, din punct de vedere meteorologic, a RAR-urilor extrase din datele radar. Astfel, propunem un nou clasificator numit $RadRAR$ (*Radar products' values prediction using Relational Association Rules*) – predicția valorilor produselor radar folosind regulile de asociere relațională) pentru prognoza pe termen scurt a furtunilor convective pe baza datelor radar.

Datele radar utilizate în experimentele noastre sunt furnizate de un radar meteorologic tip WSR-98D [NOA18]. În studiul actual ne focalizăm pe un singur produs meteorologic, și anume $R01$. Am decis să selectăm R01, deoarece este unul dintre cele mai relevante produse radar utilizate de meteorologii operaționali pentru prognoza pe pe termen scurt a vremii .

În consecință, atribuim fiecărei locații l din harta analizată (grila de date) la momentul de timp t un vector de dimensionalitate mare ale cărui elemente sunt valorile R01 pentru locațiile situate într-o vecinătate de lățime l a locației la momentul de timp t-1. Subliniem faptul că eticheta instanței de dimensionalitate l^2 descrisă anterior este valoarea lui R01 pentru punctul geografic din centrul vecinătății la momentul de timp t.

Am selectat o valoare de 13 pentru diametrul vecinătății, deoarece reprezintă aproximativ 5 kilometri în lumea fizică și această distanță determină de obicei gradienți mici ai parametrilor meteorologici. Valorile de reflectivitate peste un anumit *prag* (35 dBZ este utilizat în general [HSZ19, DW93]) sunt indicatori ai apariției unor potențiale furtuni, cu severitate de la moderată la ridicată. Astfel, împărțim setul de date D în două clase de instanțe: clasa *pozitivă* (denotată și ca „+") reprezintă instanțele etichetate cu valori R01 mai mari de 35, în timp ce clasa *negativă* (notată ca "-") reprezintă instanțele etichetate cu valori R01 mai mici sau egale cu 35. Astfel avem 2 seturi de date, D_+ și D_- pentru date pozitive și, respectiv, negative.

Am propus modelul $RadRAR$, un clasificator care este antrenat doar pe D_- și care va învăța să prezică, pe baza vecinătății unei anumite locații la momentul t, dacă valoarea ecoului radar la momentul t+1 va fi mai mare de 35 dBZ. Predicția se bazează pe estimarea probabilității p_- ca o anumită instanță 169-dimensională să aparțină clasei "-".

Procesul de clasificare pe care îl propunem are loc în două etape: *antrenarea* și *testarea*. În timpul etapei de antrenare, va fi construit un model de clasificare format dintr-un set de RAR-uri interesante din setul D_-, iar în timpul testării, modelul construit în timpul antrenării va fi aplicat pentru a decide

clasa ("+" sau "-") pentru o anumită instanță de testare.

Pentru evaluarea performanței modelului $RadRAR$, acesta este testat pe seturi de date care conțin atât instanțe *positive*, cât și *negative*, care sunt complet disjuncte de setul de date de antrenare. Pentru un set de date de testare, se calculează *matricea de confuzie* formată din patru valori: pozitive adevărate - TP, negative adevărate - TN, pozitive false - FP și negative false - FN. Ca măsuri de evaluare, folosim patru măsuri calculate pe baza valorilor din matricea de confuzie, utilizate în învățarea supervizată pentru evaluarea performanței clasificatorilor binari: *sensibilitatea* sau *probabilitatea de detecție* (**POD** = $\frac{TP}{TP+FN}$), *specificitatea* sau rata negativelor adevărate (**Spec** = $\frac{TN}{TN+FP}$), *rata alarmelor false* (**FAR** = $\frac{FP}{TP+FP}$) și *suprafața de sub curba ROC*(**AUC** = $\frac{\textbf{POD}+\textbf{Spec}}{2}$). În plus, luăm în considerare și măsura *indicele de succes critic* (**CSI** – critical success index), care este de obicei utilizată pentru prognoza pe termen scurt a furtunilor convective pe baza datelor radar - CSI = $\frac{TP}{TP+FN+FP}$. Toate măsurile de evaluare menționate anterior variază în [0, 1]. Cu excepția FAR care trebuie minimizată, valorile mai mari pentru toate celelalte măsuri de evaluare indică clasificatori mai buni.

Studiul de caz utilizat în experimentele noastre setul de date furnizate de radar în 5 iunie 2017, o zi cu instabilitate atmosferică moderată manifestată prin furtuni însoțite de ploi abundente și grindină de dimensiuni medii. În zona din regiunea centrală a Transilvaniei au existat două episoade distincte cu evenimente meteorologice intense în 5 iunie 2017. Ne restrângem la un experiment de dimensiuni reduse, deoarece scopul nostru este de a stabili o dovadă de concept a relevanței utilizării RAR-urilor pentru prognoza pe pe termen scurt a vremii. Datele utilizate pentru antrenarea modelului $RadRAR$ au fost colectate la aproximativ 14:37 UTC (în mijlocul evenimentului sever). Seturile de date D_+ și D_- colectate din datele radar brute constau din 1321 și, respectiv, 19991 instanțe.

În experimentele noastre, în procesul de extragere a regulilor, sunt luate în considerare două posibile relații între valorile atributelor: $\mathcal{R} = \{\leq, \geq\}$. După ce relațiile au fost definite, setul RAR_- al regulilor interesante de asociere relațională a fost extras din D_-.

Abordările din literatura de specialitate care sunt cele mai similare cu abordarea noastră sunt cele propuse de Yan Ji [Ji17] și Han et al. [HSZ+17, HSZ19]. Pentru a evidenția mai bine eficiența $RadRAR$ ca detector de anomalii, l-am înlocuit cu un *autoencoder* (AE). AE-urile a fost construite folosind platforma Keras din Python, în spate având platforma Tensorflow.

Analizând rezultatele, am observat că propunerea noastră, $RadRAR$, oferă rezultate mai bune pentru măsurile de evaluare în 8 din 9 comparații. Din rezultate putem concluziona că RAR-urile descoperite în din datele radar sunt eficiente pentru a prezice dacă valorile ecoului radar sunt mai mari de 35dBZ, obținând performanțe care sunt în general mai bune decât rezultatele din literatura de specialitate [HSZ19, HSZ+17].

Ca o dovadă de concept, am introdus în această secțiune un nou model de clasificare, , bazat pe descoperirea *regulilor de asociere relațională* interesante cu scopul predicției dacă valorile ecoului radar vor fi mai mari de 35dBZ. Astfel, pe baza valorilor prezise, abordarea este utilă pentru discriminarea între vreme normală și vreme furtunoasă. Pentru evaluarea performanței $RadRAR$ au fost folosite date radar reale furnizate de Administrația Națională de Meteorologie din România.

3.3 $XNow$: O tehnică de învățare profundă convoluțională pentru prognoza pe termen scurt a vremii bazată pe date radar

Am introdus în lucrarea noastră originală [SCIM20], un model de *rețele neuronale convoluționale*, $XNow$, pentru predicția pe termen scurt a datelor radar prin adaptarea arhitecturii Xception [Cho17] utilizată în principal în literatura de specialitate pentru procesarea imaginilor. Experimentele efectuate

pe date radar reale evidențiază faptul că modelul de învățare profundă propus este capabil să prezică cu exactitate valoarea datelor radar la un moment dat într-o anumită regiune geografică, pe baza valorilor lor istorice.

Datele radar exportate sunt stocate ca matrice bidimensională (grilă) în care fiecare punct corespunde unei locații geografice și conține valoarea unui produs radar la un moment dat dat. Astfel, este disponibilă o succesiune de matrici, fiecare matrice corespunzând unui anumit moment de timp t și unui anumit produs meteorologic p (de ex. R01).

Următorul pas aplicat înainte de a construi modelul de învățare profundă $XNow$ a fost aplicarea unui pas de preprocesare pe secvențele S_t pentru corectarea unor valori eronate înregistrate de radar. Pentru a evita aceste erori am decis să le înlocuim cu o estimare. Estimarea este o medie ponderată a valorilor dintr-o vecinătate (o matrice de 13 pe 13 cu punctul de estimat în centru) în care greutatea este distanța euclidiană între vecin și punct.

Funcția țintă în problema noastră de învățare este maparea f astfel încât pentru o anumită grilă de date G_t, $XNow$ va trebui să furnizeze o estimare a grilei de date 3D G_{t+1} care conține valorile produselor radar la momentul $t + 1$. Pentru realizarea învățării, modelul $XNow$ va fi obținut prin antrenarea unei arhitecturi Xception adaptate. O instanță de antrenare este sub forma (G_t, G_{t+1}).

Arhitectura Xception originală abstractizează intrarea fiecărui strat, astfel încât în final obținem o reprezentare compactă a acestuia din care se obține o singură valoare, reprezentând predicția. Dar obiectivul nostru este de a reconstrui intrarea originală, similar comportamentului unei arhitecturi codificator-decodor, păstrând în același timp eficacitatea Xception ca rețea neuronală convoluțională. În acest sens, considerăm o versiune ușor modificată a versiunii clasice, prin înlocuirea straturilor sale finale.

Vom folosi 70% din setul de date pentru antrenarea modelului $XNow$, 20% pentru validarea modelului, iar restul de 10% va fi utilizat pentru testare. Pentru evaluarea performanței $XNow$, valoarea RMSE este calculată pe datele de test descrise anterior. Datele radar utilizate în experimentele noastre conțin o mulțime de puncte de date cu valoare zero, de aceea va fi furnizată și valoarea RMSE pentru valorile diferite de zero, notat cu **RMSE**_nonzero.

Experimentele au fost efectuate folosind datele furnizate de ANM și reprezintă datele colectate de radar în 10 zile. Zilele au fost selectate astfel încât în unele zile au existat evenimente meteorologice semnificative, în timp ce în alte zile nu a existat aproape nici o activitate meteorologică, astfel încât să semene cât mai mult cu vremea tipică de vară. Datele provin de la un radar situat în centrul Transilvaniei care oferă date pentru o zonă mare.

Rezultatele au arătat că valorile RMSE sunt puțin mai mari pentru valorile diferite de zero decât pentru datele incluzând zero. Cu toate acestea, valoarea de **2.282** care e media textbf RMSE_nonzero, dacă este normalizată, devine aproximativ 3%, evidențiind o performanță foarte bună a $XNow$.

Pentru a evidenția mai bine eficiența $XNow$ (adică modelul Xception îmbunătățit), experimentele au fost efectuate folosind și arhitectura clasică Xception. S-au calculat valorile medii RMSE calculate pe mai multe antrenări ale $XNow$. Analizând rezultatele, am observat că $XNow$ oferă valori RMSE mai bune decât arhitectura clasică Xception pentru datele curățate - observăm o îmbunătățire de aproape 2 ori față de media **RMSE**_nonzero. Mai mult decât atât, *deviația standard* a valorilor RMSE după mai multe antrenări este mică și aste duce la un interval de confidență mic, evidențiind stabilitatea modelului $XNow$.

Abordarea din literatura de specialitate care este cea mai asemănătoare cu abordarea noastră este cea propusă de Yan Ji [Ji17]. Valorile RMSE exacte obținute în predicția valorilor R nu sunt furnizate, ci doar *rata de succes* definită ca procentul cazurilor în care eroarea absolută este mai mică sau egală cu 5. Minima, maximul și media valorilor *ratei de succes* sunt raportate de Yan Ji. Pornind de la acestea, am dedus limitele inferioare ale intervalului valorilor RMSE. Această limită este destul de

joasă (fiind obținută atunci când toate valorile au fost exact prezise) și, prin urmare, este greu de dedus o aproximare exactă a valorilor RMSE. Cel mai bun model $XNow$ a obținut o performanță mai bună decât rețeaua neuronală artificială propusă de Yan Ji [Ji17].

Am introdus în această secțiune un model de *rețele neuronale convoluționale*, $XNow$ pentru prezicerea, într-o manieră de învățare supăervizată, a valorilor viitoare ale produselor radar, cu scopul de a asista meteorologii în procesele de luare a deciziilor (de exemplu, furnizarea de alerte nowcasting). Au fost efectuate experimente pe date radar reale furnizate de Administrația Națională de Meteorologie a României. Pentru evidențierea eficacității modelului, $XNow$ a fost comparat cu arhitectura clasică Xception, iar rezultatele obținute au fost comparate și cu performanța actuală a soluțiilor existente prezentate în literatura de specialitate. S-a obținut o medie a *RMSE normalizată* mai mică de 3%, evidențiind o performanță foarte bună a regresorului $XNow$.

Concluzii

Scopul cercetării noastre de doctorat, conform titlului acestei teze, a fost de a dezvolta noi modele de învățare automată, atât supervizate, cât și nesupervizate, care să fie utilizate în contextul prognozei pe pe termen scurt a vremii .

Pentru partea nesupervizată a cercetării noastre, am ales modelul rețelelor neuronale cu auto-organizare (SOM) pentru studiu. Am dezvoltat două modele de date pentru datele radar și două tehnici pentru a aplica SOM-ul pe date. Primul model are la bază obiectivul de a descoperi modul în care produsele radar evoluează de-a lungul scanărilor radar consecutive. Prin interpretarea U-matricilor rezultate, am arătat că valorile produselor radar se schimbă lent în timp, cu excepția unor momente specifice legate de fenomene meteorologice severe. Al doilea model de date are la bază obiectivul studierii relației dintre valoarea unui produs radar într-o locație la un moment dat și valorile produselor radar dintr-o vecinătate a acelei locații în momentele de timp anterioare. Interpretând rezultatele SOM-ului cu acest model de date, am arătat că pentru valori similare ale unui produs, vecinătățile din momentele anterioare sunt similare. De asemenea, am creat o măsură de evaluare - eroarea medie de similaritate - care arată că rezultatele experimentelor noastre SOM sunt semnificative.

Cercetările noastre privind partea de învățare supervizată a proiectului au culminat cu dezvoltarea a trei noi modele de învățare automată pentru prognoza pe pe termen scurt a vremii: *NowDeepN*, *RadRAR* și *XNow*. Am dezvoltat *NowDeepN*, un ansamblu de 13 rețele neuronale, fiecare prezicând un produs radar diferit la o anumită locație pe baza tuturor produselor radar dintr-o vecinătate a acelei locații. Am arătat că *NowdeepN* funcționează destul de bine, în comparație cu alte modele din literatură, comparația fiind favorabilă modelului nostru în 5 din 7 cazuri. *RadRAR* are scopul de a clasifica dacă valoarea unui produs radar va fi peste sau sub un prag. textit RadRAR învață mai întâi regulile din date, separat între cele două clase, apoi, pe baza regulilor extrase, poate prezice dacă valoarea R01 la o locație este peste sau sub pragul de 35 dBZ, pe baza valorilor lui R01 într-o vecinătate a acelei locații la pasul anterior. Am arătat că *RadRAR* este destul de performant pentru aceast lucru, comparându-se favorabil cu alte modele din literatura de specialitate. Ultimul model de învățare automată supervizată pe care l-am dezvoltat a fost *XNow*. *XNow* este capabil să prezică toate datele pentru o singură etapă, pe baza datelor din pasul anterior. Am demonstrat empiric că modelul are rezultate foarte bune, depășind ușor celelalte modele din literatură.

În viitor ne propunem să continuăm dezvoltarea acestor modele. Pentru *RadAR* ne gândim să îmbunătățim algoritmul de extragere a regulilor și să optimizăm datele din care să extragă regulile și, de asemenea, numărul de reguli.

Pentru continuarea eforturilor noastre pentru crearea unor modele mai bune de învățare automată pentru prognoza pe pe termen scurt a vremii, focalizarea noastră va fi pe modelul *XNow*, deoarece a avut cele mai bune rezultate dintre cele trei modele și, de asemenea, este cel mai scalabil. Deci, în viitoarele proiecte de cercetare, intenționăm să extindem modelul XNow pentru a putea prezice mai mult de un pas de timp în viitor. De asemenea, ne gândim să folosim mai mulți pași de timp anteriori și să creștem cantitatea de date de antrenament de la zile la săptămâni sau luni.

Bibliografie

[CAG19] Liana-Maria Crivei, Mihai Andrei, and Czibula Gabriela. A study on applying relational association rule mining based classification for predicting the academic performance of students. In *KSEM 2019 : The 12th International Conference on Knowledge Science, Engineering and Management, LNAI 11775*, pages 287–300, 2019.

[CBC12] Gabriela Czibula, Maria-Iuliana Bocicor, and Istvan Gergely Czibula. Promoter sequences prediction using relational association rule mining. *Evolutionary Bioinformatics*, 8:181–196, 04 2012.

[Cho17] F. Chollet. Xception: Deep Learning with Depthwise Separable Convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1800–1807, 2017.

[CMA+21] Gabriela Czibula, Andrei Mihai, Alexandra-Ioana ALbu, Istvan Czibula, Sorin Burcea, and Abdelkader Mezghani. $autonowp$: An approach using deep autoencoders for precipitation nowcasting based on radar echo prediction. *Mathematics*, in press, 2021.

[CMC19a] G. Czibula, A. Mihai, and L.M. Crivei. A novel relational association rule mining classification model applied for academic performance prediction. page accepted for publication. Elsevier, 2019.

[CMC19b] G. Czibula, A. Mihai, and I.G. Czibula. Radrar: A relational association rule mining approach for nowcasting based on predicting radar products' values. pages 300–309. Elsevier, 2019.

[CMt21] Gabriela Czibula, Andrei Mihai, and Eugen Mihuleţ. $nowdeepn$: An ensamble of deep learning models for weather nowcasting based on radar products' values prediction. *Applied Sciences*, 11:121, 2021.

[CMtT19] Gabriela Czibula, Andrei Mihai, Eugen Mihuleţ, and Daniel Teodorovici. Using self-organizing maps for unsupervised analysis of radar data for nowcasting purposes. *Procedia Computer Science*, 159:48–57, 2019. 23rd International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2019).

[CvMG+14] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics.

[DHG+14] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. *CoRR*, abs/1411.4389, 2014.

[DW93] M. Dixon and G. Wiener. TITAN: Thunderstorm Identification, Tracking, Analysis, and Nowcasting - A radar-based methodology. *Journal of Atmospheric and Oceanic Technology*, 10(6):785–797, 1993.

[GWD14] Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. *CoRR*, abs/1410.5401, 2014.

[HK06] Rob J. Hyndman and Anne B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679 – 688, 2006.

[HS97] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November 1997.

[HS13] Michiel Hermans and Benjamin Schrauwen. Training and analysing deep recurrent neural networks. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 190–198. Curran Associates, Inc., 2013.

[HSZ+17] Lei Han, Juanzhen Sun, Wei Zhang, Yuanyuan Xiu, Hailei Feng, and Yinjing Lin. A machine learning nowcasting method based on real-time reanalysis data. *Journal of Geophysical Research: Atmospheres*, 122(7):4038–4051, 2017.

[HSZ19] Lei Han, Juanzhen Sun, and Wei Zhang. Convolutional Neural Network for Convective Storm Nowcasting Using 3D Doppler Weather Radar Data. *arXiv e-prints*, page arXiv:1911.06185, Nov 2019.

[Ji17] Yan Ji. Short-term precipitation prediction based on a neural network. In *3rd International Conference on Artificial Intelligence and Industrial Engineering*, AIIE 2017, pages 246–251, 2017.

[Ker18] Keras. The Python Deep Learning library, 2018. https://keras.io/.

[KK96] S. Kaski and T. Kohonen. Exploratory data analysis by the self-organizing map: Structures of welfare and poverty in the world. In *Neural Networks in Financial Engineering. Proceedings of the Third International Conference on Neural Networks in the Capital Markets*, pages 498–507. World Scientific, 1996.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pages 1097–1105. Curran Associates Inc., 2012.

[KTO+07] Peter K. Kihato, Heizo Tokutaka, Masaaki Ohkita, Kikuo Fujimura, Kazuhiko Kotani, Yoichi Kurozawa, and Yoshio Maniwa. Spherical and torus som approaches to metabolic syndrome evaluation. In Masumi Ishikawa, Kenji Doya, Hiroyuki Miyamoto, and Takeshi Yamakawa, editors, *ICONIP (2)*, volume 4985 of *Lecture Notes in Computer Science*, pages 274–284. Springer, 2007.

[Lip15] Zachary Chase Lipton. A critical review of recurrent neural networks for sequence learning. *CoRR*, abs/1506.00019, 2015.

[LO92] J. Lampinen and E. Oja. Clustering properties of hierarchical self-organizing maps. *Journal of Mathematical Imaging and Vision*, 2(3):261–272, 1992.

[MCt19] Andrei Mihai, Gabriela Czibula, and Eugen Mihuleţ. Analyzing meteorological data using unsupervised learning techniques. In *ICCP 2019: IEEE 15th International Conference on Intelligent Computer Communication and Processing*, pages 1–8. IEEE Computer Society, 2019.

[Mih20] Andrei Mihai. Using self-organizing maps as unsupervised learning models for meteorological data mining. In *IEEE 14th International Symposium on Applied Computational Intelligence and Informatics (SACI 2020)*, pages 23–28. IEEE Hungary Section, 2020.

[Mit97] Thomas M. Mitchell. *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition, 1997.

[NOA18] NOAA's National Weather Service. Radar Operations Center. NEXRAD Technical Information, 2018. https://www.roc.noaa.gov/WSR88D/Engineering/NEXRADTechInfo.aspx.

[RFB15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[SCC06] Gabriela Serban, Alina Campan, and Istvan Gergely Czibula. A programming interface for finding relational association rules. *International Journal of Computers, Communications & Control*, I(S.):439–444, June 2006.

[SCIM20] Ioana Angela Socaci, Gabriela Czibula, Vlad-Sebastian Ionescu, and Andrei Mihai. $xnow$: A deep learning technique for nowcasting based on radar products' values prediction. In *IEEE 14th International Symposium on Applied Computational Intelligence and Informatics (SACI 2020)*, pages 117–122. IEEE Hungary Section, 2020.

[SCW+15] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 802–810, Cambridge, MA, USA, 2015. MIT Press.

[SHK+14] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[SK99] Panu Somervuo and Teuvo Kohonen. Self-organizing maps and learning vector quantization for feature sequences. *Neural Processing Letters*, 10:151–159, 1999.

[Swe18] Swedish Meteorological and Hydrological Institute. Cooperation is a must for adaptation to and mitigation of climate change, 2018. https://www.smhi.se/en/news-archive/.

[TS19] Quang-Khai Tran and Sa-kwang Song. Computer vision in precipitation nowcasting: Applying image quality assessment metrics for training deep neural networks. *Atmosphere*, 10(5):244, May 2019.

[TSM21] Kevin Trebing, Tomasz Staczyk, and Siamak Mehrkanoon. Smaat-unet: Precipitation nowcasting using a small attention-unet architecture. *Pattern Recognition Letters*, 145:178–186, 2021.

[WKS16] Michael Wand, Jan Koutník, and Jürgen Schmidhuber. Lipreading with long short-term memory. *CoRR*, abs/1601.08188, 2016.

[YJM+20] Qing Yan, Fuxin Ji, Kaichao Miao, Qi Wu, Yi Xia, and Teng Li. Convolutional residual-attention: A deep learning approach for precipitation nowcasting. *Advances in Meteorology*, 2020:6484812, Feb 2020.