UNIVERSITATEA BABEŞ-BOLYAI
FACULTATEA DE MATEMATICĂ ȘI INFORMATICĂ

Dezvoltarea unor modele de învățare profundă pentru probleme complexe

Rezumatul tezei de doctorat

Cuvinte cheie: instruire automată, bioinformatică, vedere artificială, instruire profundă

Doctorand: Mihai Teletin
Conducător științific: Prof. Dr. Czibula Gabriela

Iulie 2020

Cuprins

Cuprinsul tezei

Lista publicațiilor

Toate clasamentele sunt listate în conformitate cu clasificarea jurnalelor [1] și conferințelor [2] în informatică.

Publicații în Web of Science - Science Citation Index Expanded

[CCT19] Gabriela Czibula, Carmina Codre, and **Mihai Teletin**. *AnomalP: A new approach for detecting anomalous protein conformations using deep autoencoders.* Expert Systems with Applications, 2019, under review (**IF=4.292**)

[ATC18] Silvana Albert, **Mihai Teletin**, and Gabriela Czibula. *Analysing protein data using unsupervised learning techniques.* International Journal of Innovative Computing, Information and Control (IJICIC), pp. 861-880, Volume 14, Number 3, June 2018. (**IF=1.667**)

Categoria C, 2 puncte.

Publicații in Web of Science - Conference Proceedings Citation Index

[TC20] **Mihai Teletin** and Gabriela Czibula. *CVSimP: An approach for predicting proteins' structural similarity using one-shot learning.* IEEE 14th International Symposium on Applied Computational Intelligence and Informatics, SACI 2020, Timisoara, in press

Categoria C, 2 puncte.

[DT19] Lorand Dobai, **Mihai Teletin**. *A document detection technique using convolutional neural networks for optical character recognition systems.*European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Bruges, Belgium, ESANN, pp. 547–552, 2019.

Categoria B, 4 puncte.

[TCC19] **Mihai Teletin**, Gabriela Czibula and Carmina Codre. *AutoSimP: An approach for predicting proteins' structural similarities using an ensemble of deep autoencoders.*The 12th International Conference on Knowledge Science, Engineering and Management (KSEM 2019), LNAI 11776, pp. 49–54, 2019.

Categoria B, 4 puncte.

[TCAB18] **Mihai Teletin**, Gabriela Czibula, Silvana Albert, and Mariana-Iuliana Bocicor. *Using unsupervised learning methods for enhancing protein structure insight.* International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES 2018, Belgrade, Serbia, Procedia Computer Science, 126, pp. 126-135, 2018.

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

Categoria B, 2 puncte.

[TCB+18] **Mihai Teletin**, Gabriela Czibula, Mariana-Iuliana Bocicor, Silvana Albert, and Alessandro Pandini. *Deep autoencoders for additional insight into protein dynamics.* International Conference on Artificial Neural Networks (ICANN), Rhodes, Greece, LNCS, volume 11140, pp. 78-89, 2018.

Categoria B, 1.33 puncte.

[TCB19] **Mihai Teletin**, Gabriela Czibula and Maria-Iuliana Bocicor. *Using clustering models for uncovering proteins' structural similarity.* IEEE 13th International Symposium on Applied Computational Intelligence and Informatics, SACI 2019, Timisoara, Romania, pp. 185-190, 2019

Categoria C, 2 puncte.

[TD19] **Mihai Teletin**, Lorand Dobai. *Lightweight models for fruits recognition.* IEEE 13th International Symposium on Applied Computational Intelligence and Informatics, SACI 2019, Timisoara, Romania, pp. 69-74, 2019.

Categoria C, 2 puncte.

[ACT18] Silvana Albert, Gabriela Czibula, and **Mihai Teletin**. *Analyzing the impact of protein representation on mining structural patterns from protein data.* IEEE 12th International Symposium on Applied Computational Intelligence and Informatics, SACI 2018, Timisoara, Romania, pp. 533-538, 2018.

Categoria C, 2 puncte.

[ICT17] Vlad-Sebastian Ionescu, Gabriela Czibula, **Mihai Teletin**. *Supervised learning techniques for body mass estimation in bioarchaeology.* IEEE 7th International Workshop on Soft Computing Applications (SOFA), Springer, pp. 71-86, 2017.

Categoria C, 2 puncte.

[ITV16] Vlad-Sebastian Ionescu, **Mihai Teletin**, Estera-Maria Voiculescu. *Machine learning techniques for age at death estimation from long bone lengths.* IEEE 11th International Symposium on Applied Computational Intelligence and Informatics (SACI 2016), Timisoara, Romania, pp. 457 - 462, 2016.

Categoria C, 2 puncte.

Publications în jurnale și conferințe indexate in baze de date internaționale

[Tel17] **Mihai Teletin**. *Machine Learning Techniques for Detecting False Signatures.* Studia Universitatis Babes-Bolyai, Informatica 62(1), pp. 49–59, 2017. (**indexed Mathematical Reviews**)

Categoria D, 1 point.

[BPC+17] Maria-Iuliana Bocicor, Alessandro Pandini, Gabriela Czibula, Silvana Albert, and **Mihai Teletin**. *Using computational intelligence models for additional insight into protein structure.* Studia Universitatis Babes-Bolyai, Informatica 62(1), pp. 107–119, 2017. (**indexed Mathematical Reviews**)

Categoria D, 0.33 puncte.

Scorul publicațiilor: 26.66 puncte.

Introducere

Principalul domeniu de studiu urmărit în teza noastră de doctorat este instruirea automată (eng. Machine Learning (*ML*)). Teza de doctorat este intitulată "Contribuții în dezvoltarea de modele de învățare profundă pentru probleme complexe". Cercetarea noastră este axată pe dezvoltarea de noi modele deep learning pentru rezolvarea problemelor complexe din două domenii, respectiv *bioinformatică* și vedere aritificială (eng. *computer vision*).

Învățarea automată este principala direcție de cercetare în domeniul Inteligenței artificiale (AI). Scopul acestei discipline este de a dezvolta modele care fac predicții și care sunt capabile să îmbunătățească aceste predicții acumulând experiență. În prezent, aplicabilitatea ML în vederea artificială și bioinformatică este un subiect foarte popular. Prin urmare, o parte semnificativă a cercetării din aceste două domenii particulare este orientată către ML.

Cercetările noastre în *ML* au început cu dezvoltarea de modele neuronale artificiale pentru două sarcini importante în arheologia computațională: predicția *masei corporale* și a *vârstei la deces* estimate din resturile scheletice umane. Ambele probleme au o importanță majoră în cercetarea paleontologică și arheologică, deoarece pot furniza informații utile despre populațiile din trecut, precum sănătatea lor, aspecte sociale diferite, influența factorilor de mediu și altele. Sarcina de a estima masa corporală din resturile scheletice umane pe baza măsurătorilor osoase a fost investigată dintr-o perspectivă *ML* în [ICT17]. Au fost propuse două modele de regresie bazate pe învățare supervizată, folosind *rețele neuronale artificiale* și *mașini cu suport vectorial*, pentru a exprima bune mapări neliniare între măsurători scheletice și masa corporală. Mai multe experimente efectuate pe un set de date disponibil public au arătat că aplicațiile propuse de algoritmi bazaţi pe învățare automată duc la rezultate mai bune decât cele existente în literatură. În [ITV16] am aplicat *rețele neuronale artificiale* și *mașini cu suport vectorial* pentru estimarea vârstei la deces și am arătat că acestea depășesc abordările matematice existente, pe o serie de studii de caz derivate din datele disponibile public.

O subdisciplină importantă a *ML* care s-a dezvoltat din studiul rețelelor neuronale artificiale este deep learning. Domeniul a înregistrat un succes notabil în diferite domenii, reușind să îmbunătățească performanțele recente. Potențialul metodelor de învățare automată a fost utilizat cu succes pentru rezolvarea problemelor care au fost considerate foarte dificile. De exemplu, una dintre cele mai complexe probleme din domeniul vederii artificiale, problema ImageNet [DDS+09] a fost rezolvată folosind o rețea neuronală convoluțională [KSH12, SVI+15, Cho16]. Pe de altă parte, o problemă complexă în domeniul bioinformaticii este analiza proteinelor. Diferite lucrări au subliniat faptul că tehnicile *ML* sunt potrivite pentru soluționarea acestei probleme [LJN06, BDF+15].

Probleme abordate. Motivaţie

Teza este axată pe dezvoltarea de soluții *ML* pentru două domenii distincte din viața reală, și anume *computer vision* și *bioinformatică*. La prima vedere, aceste două domenii

par a fi diferite. Cu toate acestea, se poate susține că ambele includ probleme dificile și greu de rezolvat. Mai mult, dintr-o perspectivă computațională, principala legătură dintre problemele din viața reală din aceste două domenii de studiu este *complexitatea* lor.

Complexitatea problemelor abordate în teza curentă poate fi înțeleasă din mai multe puncte de vedere. În primul rând, o parte a complexității unei probleme poate fi legată de *dificultate*, ceea ce face ca problema să nu poată fi rezolvată prin programare tradițională, adică algoritmi și metode clasice. Mai mult decât atât, considerăm *complexă* o problemă legată și de datele de intrare care trebuie prelucrate: în majoritatea cazurilor, o mare cantitate de date (eventual neetichetate) trebuie analizată; în anumite cazuri, datele conțin "zgomot" (eng. *noise*). Din aceste motive, considerăm metodele *ML* foarte potrivite pentru a gestiona *complexitatea* problemelor, așa cum am discutat anterior. Pe de altă parte, dintr-o perspectivă computațională, ne referim la complexitatea unei probleme din punct de vedere al *NP-completitudinii*. Problemele NP-complete sunt probleme ale căror soluții pot fi verificate în timp polinomia. În general, ambele clase de probleme nu pot fi rezolvate în timp polinomial [BCB94]. Din perspectiva vederii computerizate, există câteva rezultate din literatura de specialitate care arată că unele probleme de bază, cum ar fi potrivirea imaginii sunt *NP-complete* [KU03]. Pe domeniul *bioinformaticii*, există câteva rezultate preliminare care sugerează că plierea proteinei este o problemă *NP-completă* [BL98, BBCG13].

Concentrându-ne mai mult asupra domeniilor noastre de cercetare, putem argumenta că este foarte greu să proiectăm o soluție pentru problemele din *bioinformatică* și *computer vision* folosind algoritmi clasici. Mai mult, este de asemenea cunoscut faptul că meta și hipereuristicile nu pot atinge performanțe de ultimă generație. Pe de altă parte, capacitatea tehnicilor *ML* de a se adapta la situații necunoscute prin învățare poate fi utilă. De fapt, această capacitate ne motivează să folosim astfel de tehnici pentru abordarea celor două domenii de studiu, propunând noi tehnici care sunt capabile să atingă performanța de ultimă oră.

Ne propunem să ne concentrăm cercetarea către dezvoltarea de modele *ML* pentru sarcini de analiză a proteinelor și *computer vision*. După cum s-a arătat anterior, principala legătură dintre cele două domenii pe care le studiem este complexitatea lor. Considerăm că acesta este motivul principal pentru care tehnicile *deep learning* pot fi foarte eficiente pentru rezolvarea problemelor legate de aceste domenii.

Analiza proteinelor prezintă un interes deosebit atât în cercetarea biologiei computaționale, cât și în domeniul bioinformaticii. Proteinele sunt molecule mari, complexe, cu roluri esențiale în funcționarea organismelor vii. Înțelegerea mecanismelor de bază prin care proteinele își realizează structurile și substructurile, precum și cei implicați în tranzițiile conformaționale pot contribui la o înțelegere mai profundă a proceselor biologice implicate. Deși structura 3D stabilă a unei proteine este definită printr-o topologie unică, această structură nu este statică și acum este acceptat pe scară largă faptul că proteinele sunt obiecte dinamice [TT09]. Conform unor diverși factori externi din mediul proteinei (de exemplu, temperatura, interacțiunea cu alte molecule), modificările structurilor proteinelor apar în timpul funcțiilor lor biologice. Astfel, o proteină va dobândi un număr limitat de conformații și va avea capacitatea de a tranziționa între conformații alternative. Înțelegerea dinamicii proteinelor și modul în care se produc aceste tranziții conformaționale este esențială pentru înțelegerea interacțiunilor biomoleculare, ceea ce este de o importanță crucială în procesul dezvoltării de noi medicamente care pot inhiba comportamentul necontrolat al proteinelor [MMC08].

Atât importanța, cât și complexitatea problemei ne motivează să investigăm utilitatea modelelor și metodelor *ML* pentru analiza și detectarea modificărilor conformaționale ale proteinelor. Datorită vitezei și complexității dinamicii proteinelor, analiza unei proteine pentru o perioadă relativ scurtă de timp poate fi utilizată pentru a extrage un volum mare de date. Este cunoscut faptul că o astfel de cantitate masivă de date poate fi exploatată cu

succes folosind diverse tehnici de *ML*.

În domeniul *analizei proteinei* vom concentra cercetările noastre spre dezvoltarea metodologiilor *ML* pentru analiza structurii proteinelor cu scopul de a ne ajuta să înțelegem și să analizăm complexitatea proteinelor: să utilizăm tehnici de învățare nesupervizată pentru analiza structurii proteice inter și intra; dezvolta tehnici eficiente de codificare; folosind autoencodere pentru clasificarea proteinelor în superfamilii și detectarea conformațiilor proteice anomale.

În al doilea rând, ne propunem să dezvoltăm tehnici *deep learning* pentru unele probleme *computer vision*: îmbunătățirea performanței pentru o sarcină de clasificare, și anume clasificarea semnăturilor; să demonstrăm capacitatea rețelelor neuronale convoluționale pentru preprocesarea imaginilor prin dezvoltarea de tehnici de localizare a documentelor și îndrptarea acestora.

Structura tezei

Restul tezei este structurat după cum urmează. Prima direcție de cercetare legată de bioinformatică este analiza intra proteinelor și va fi abordată în capitolul 1. Experimentele efectuate sunt ilustrate în Secțiunile 1.2 și 1.1. Mai exact, un experiment bazat pe *clustering* va fi descris în Secțiunea 1.1. Din perspectiva metodologiei, prezentăm tehnicile de clustering utilizate, măsurile de evaluare, unele tehnici de vizualizare și procesul de codificare a proteinelor din Secțiunea 1.1.1. Mai mult, rezultatele și discuțiile experimentale sunt prezentate în Secțiunea 1.1.3. Un experiment bazat pe autoencodere este prezentat în 1.2. Metodologia este introdusă în Secțiunea 1.2.1 și acoperă detaliile reprezentării proteinelor și setările parametrilor autoencoderului. Experimentele și rezultatele obținute sunt discutate în Secțiunea 1.2.2.

Capitolul 2 examinează rezultatele noastre legate de analiza inter proteine. Prin urmare, Secțiunea 2.1 conține un experiment bazat pe clustering care reușește să obțină prformanțe mai bune decât lucrările anterioare pe același subiect. Trei reprezentări pentru tranzițiile conformaționale ale proteinelor sunt propuse în Secțiunea 2.1.1. Experimentele sunt efectuate folosind toate aceste reprezentări, iar rezultatele obținute sunt descrise și comparate în Secțiunea 2.1.3. Metoda este apoi extinsă dintr-o perspectivă diferită în Secțiunea 2.2. Mai exact, un ansamblu de autoencodere este dezvoltat pentru a rezolva aceeași problemă. Din perspectiva abordării, descriem reprezentarea datelor și modelul din Secțiunea 2.2.1. Rezultatele experimentale sunt apoi evidențiate în Secțiunea 2.2.2. Metoda prezentată în această secțiune a reușit să atingă performanța de ultimă oră. O abordare numită *AnomalP* bazată pe *deep autoencoders* pentru detectarea tranzițiilor anomale conformaționale ale proteinelor este introdusă în Secțiunea 2.3. Metoda se bazează pe un ansamblu de autoencodere și este descrisă în Secțiunea 2.3.1. Experimentele și rezultatele experimentale obținute sunt prezentate în secțiunea 2.3.2, în timp ce Secțiunea 2.3.3 oferă o analiză a rezultatelor, precum și o comparație cu lucrările similare existente. Secțiunea 2.3.4 rezumă concluziile subsecțiunii și indică direcții pentru îmbunătățiri și extinderi suplimentare.

În cele din urmă, în Capitolul 3 discutăm trei activități practice de *computer vision*. În primul rând, o metodă de verificare a semnăturii este introdusă în Secțiunea 3.1. Apoi, o soluție pentru un set de date open source pentru recunoașterea fructelor este prezentată în Secțiunea 3.2. Metoda propusă în secțiunea 3.2.1 se bazează pe modele ușoare *deep learning* și realizează performanțe de ultimă generație atât pentru timpul de execuție, cât și pentru performanța de clasificare, așa cum este descris în 3.2.2. În cele din urmă, o tehnică de localizare bazată pe *rețele neuronale convoluționale* va fi introdusă în Secțiunea 3.3. Studiem problema localizării în timp ce încercăm să îmbunătățim performanțele de ultimă generație pentru o sarcină practică, detectarea și îndreptarea documentelor. O parte importantă a unei sarcini de recunoaștere optică a caracterelor este etapa de preprocesare,

al cărei scop este de a îmbunătăți condițiile în care extragerea ulterioară a textului este efectuată. Rezultatele prezentate în Secțiunea 3.3.2 depășesc prformanța descrisă în lucrările existente și pot fi utilizate eficient pentru îmbunătățirea preciziei de recunoaștere optică a caracterelor.

Contribuții originale

Această teză este axată pe două direcții principale de cercetare, viziunea computerului și analiza proteinelor. Prin urmare, contribuțiile originale ale acestei teze sunt duble.

1. Dintr-o perspectivă bioinformatică, am proiectat și implementat noi metode bazate pe *ML* pentru efectuarea analizei proteice.
2. Din perspectiva viziunii computerului am dezvoltat soluții bazate pe *deep learning* pentru unele activități de procesare a imaginilor.

Contribuțiile legate de bioinformatică au fost incluse în capitolele 1 și 2 și au fost publicate în șapte lucrări de cercetare [BPC+17, ATC18, ACT18, TCB+18, TCAB18, TCC19, TCB19]:

- Am ilustrat experimental că informațiile obținute prin analizarea tranzițiilor conformaționale ale proteinelor captează relațiile dintre proteinele înrudite, relații care sunt confirmate din perspectivă biologică [BPC+17]. Am investigat o nouă perspectivă *ML* pentru analiza tranzițiilor conformaționale proteice și am propus o nouă formalizare a problemei discutate. Acest studiu a reprezentat punctul de plecare al cercetării noastre.
- Am investigat utilitatea *rețelelor cu auto-organizare* și a celor *fuzzy* în identificarea relației structurale dintre proteine [ATC18]. Experimentele pe care le-am efectuat pe mai multe seturi de date proteice ilustrează eficiența modelelor de învățare nesupervizate pentru a surprinde similitudinea dintre proteine. Rezultatele raportate au relevat, de asemenea, că modelele *fuzzy* sunt capabile să crească performanța modelului nesupervizat.
- S-a analizat impactul valorilor *RSA* în utilizrea *rețelelor cu autoorganizare* pentru studierea structurii interne a proteinelor [ACT18]. Folosind două reprezentări distincte, au fost efectuate două studii de caz pentru a sublinia eficiența abordării bazate pe *rețele cu autoorganizare*.
- Am studiat capacitatea autoencoderelor de a păstra și ilustra informațiile proteice în timp ce am analizat datele proteice [TCB+18]. Modelele *deep learning* au fost explorate pentru a evidenția capacitatea lor de a învăța tipare biologice relevante, cum ar fi caracteristicile structurale. Studiul și-a propus să ofere o mai bună înțelegere a modului în care evoluția tranzițiilor conformaționale proteice evoluează în timp, în cadrul mai larg al detectării automate a plierii proteinelor.
- Capacitatea clusteringului ca metodă de clasificare nesupervizată a fost demonstrată pentru investigarea relevanței valorilor *RSA* pentru a prezice tranzițiile interne ale proteinelor [TCAB18]. Cu scopul principal al studierii evoluției valorilor *RSA* între tranzițiile conformaționale, am arătat experimental că valorile *RSA* se modifică lent, pe măsură ce proteina suferă modificări conformaționale.
- Am extins abordarea de clustering pentru analiza inter-proteină [TCB19]. Am investigat trei reprezentări pentru o proteină bazată pe distribuțiile de probabilitate ale anumitor elemente structurale în cadrul tranzițiilor conformaționale și am aplicat

metode de clustering pentru a clasifica fără supervizare proteine pe baza asemănării lor structurale. Experimentele au fost efectuate pe două seturi de date proteice. Rezultatele comparative au relevat faptul că, în multe cazuri, propunerea noastră este mai performntă decât lucrări anterioară în acest subiect.

- A fost dezvoltat un sistem bazat pe un ansamblu de rețele neuronale cu autosupervizare pentru clasificarea proteinelor [TCC19]. Scopul sistemului este predicția superfamiliei unei anumite proteine, având în vedere clasa de similaritate prezisă pentru tranzițiile conformaționale ale acesteia. Experimentele au fost efectuate pe date proteice reale și au relevat eficacitatea propunerii noastre în comparație cu abordările existente similare.

Contribuțiile legate de *computer vision* sunt incluse în capitolul 3 și au fost introduse în trei lucrări de cercetare [Tel17, TD19, DT19]:

- Am aplicat metode *deep learning* pentru determinarea autenticității semnăturii [Tel17]. A fost investigată posibilitatea de a utiliza tehnici de învățare supervizate pentru a construi modele capabile să efectueze cu exactitate o astfel de analiză. Rezultatele raportate în faza de testare au fost încurajatoare pentru continuarea lucrărilor.

- A fost proiectată o metodă de detectare și îndreptare a documentelor. S-a bazat pe o rețea neuronală convoluțională și transformarea perspectivei [DT19]. Cercetarea noastră și-a propus să servească ca o etapă de preprocesare pentru un sistem de recunoaștere optică. Principala provocare a fost îmbunătățirea performanței, în special pe cadre care au fost înclinate (ușor rotite) sau care au avut fundaluri înghesuite. Metoda propusă a obținut rezultate bune de detectare a documentelor și de îndreptare e a unui set de date cu fotografii de bonuri de casă.

- S-a propus o soluție de ultimă generație pentru un set de date de clasificare a fructelor din imagini [TD19]. Soluția s-a bazat exclusiv pe rețele neuronale profunde ușoare. Am obținut performanțe de ultimă generație atât în ceea ce privește performanța de clasificare cât și viteza de execuție.

Perspectivele noastre asupra problemelor abordate sunt noi, în conformitate cu cunoștințele noastre, nu au fost încă cercetate în literatură. Suntem siguri că soluțiile bazate pe *ML* sunt aplicabile în domeniile abordate în teză, *bioinformatică* și *computer vision* și pot duce la informații interesante și valoroase, datorită capacității acestor modele de descoperi tipare ascunse în date.

***

Autorul tezei mulțumește lectorului Alessandro Pandini de la Universitatea Brunel din Londra pentru furnizarea seturilor de date proteice utilizate în experimentele care sunt descrise în capitolele 1 și 2.

Capitolul 1

Modele noi de învățare automată pentru analiză intra-proteină

Proteinele sunt elementele de construcție ale oricărui organism viu. Sunt macromolecule complexe care contribuie la menținerea mediilor celulare și au astfel roluri fundamentale în procesele biologice. Ele reprezintă rezultatul final al procesului de decodare *ADN*. Cu toate acestea, înțelegerea genezei și plierea lor este încă o piesă lipsă din literatura de biologie. În acest capitol prezentăm investigația noastră privind aplicarea metodelor *învățare nesupervizată* pentru analiza structurii proteice interne și a tranzițiilor la care participă. Scopul principal este extragerea de informații semnificative despre asemănarea structurală a proteinelor. Experimentele efectuate pe diferite seturi de date proteice subliniază eficacitatea modelelor de învățare nesupervizate pentru captarea asemănării dintre structurile proteinelor.

În acest capitol evidențiem două contribuții principale principale care au fost publicate în lucrări originale [TCAB18, TCB+18]:

- În secțiunea 1.1 folosim *clustering* ca metodă de clasificare nesupervizată pentru a studia relevanța valorilor *accesibilitate relativă a solventului* (*RSA*) pentru a analiza transformarile interne ale proteinelor [TCAB18]. Cu scopul principal de a studia evoluția valorilor *RSA* între tranzițiile conformaționale, arătăm experimental că valorile *RSA* se modifică lent, pe măsură ce proteina suferă modificări conformaționale. Studiul urmărește să ofere o mai bună înțelegere a modului în care evoluția tranzițiilor conformaționale ale proteinelor evoluează în timp, cu scopul mai larg de a înțelege mai bine dinamica internă a proteinelor.

- În secțiunea 1.2, investigăm utilitatea metodelor *ML* nesupervizate pentru descoperirea informațiilor relevante despre dinamica funcțională a proteinei [TCB+18]. *Autoencoderele* sunt explorate pentru a evidenția capacitatea lor de a învăța tipare biologice relevante, cum ar fi caracteristicile structurale. Acest studiu urmărește să ofere o mai bună înțelegere a modului în care tranzițiile conformației proteice evoluează în timp, în cadrul mai larg al detectării automate a mișcărilor funcționale.

Secțiunea 1.1 este structurată după cum urmează. Punem în evidență tehnicile de clustering utilizate, măsurile de evaluare, unele tehnici de vizualizare și procesul de codificare a proteinelor din secțiunea 1.1.1. Rezultatele experimentelor și unele comparații sunt prezentate în secțiunea 1.1.3. În cele din urmă, concluziile sunt trase în secțiunea 1.1.4.

Secțiunea 1.2 este structurată după cum urmează. Metodologia abordării este discutată în secțiunea 1.2.1 și acoperă detalii despre reprezentarea proteinelor și setările parametrilor modelului. Apoi, experimentele și rezultatele obținute sunt discutate în secțiunea 1.2.2. În cele din urmă, prezentăm pe scurt concluziile direcției de cercetare în secțiunea 1.2.3.

1.1 Folosirea de metode de învățare nesupervizată pentru analiza structurii proteinelor

În secțiunea curentă, ilustrăm diverse experimente bazate pe clustering pentru analiza tranzițiilor interne ale proteinelor. Experimentele descrise în această secțiune au fost introduse în lucrarea originală [TCAB18].

În acest capitol cercetăm utilitatea modelelor de *clustering* pentru a descoperi în mod nesupervizat informații relevante care ar oferi o mai bună înțelegere a structurii proteinelor, cu scopul mai larg de a învăța să prezicem cum evoluează proteinele în timp. În studiul nostru folosim o reprezentare internă pentru o proteină bazată pe valorile reprezentând *accesibilitate relativă a solventului* (RSA) reziduurilor de aminoacizi, având în vedere un set substanțial de tranziții conformaționale.

Prin mai multe experimente efectuate pe *patru* proteine ne propunem să obținem o dovadă empirică potrivit căreia (1) există anumite modele în care o proteină trece de la o conformație la alta și că (2) conformațiile proteice care sunt apropiate temporal sunt similare unul altuia. Experimentele noastre arată că valorile *RSA* se schimbă încet între tranzițiile proteinelor. În conformitate cu cunoștințele noastre, un studiu similar cu al nostru nu a fost efectuat până în prezent pentru analiza intra-proteină.

1.1.1 Metodologie

Metodologia experimentală utilizată pentru evaluarea eficacității clusteringului în analiza tranzițiilor conformaționale ale proteinelor este detaliată în continuare. Realizăm un experiment de clustering pentru a afla cât de bine sunt grupate conformațiile unei proteine (reprezentate ca vectori *RSA*). Mai mult, vom folosi analiza *PCA* pentru a crea vizualizarea bidimensională a datelor noastre. Implementarea scikit-learn [PVG+11] a fost folosită atât pentru clustering, cât și pentru PCA.

1.1.2 Seturi de date proteice

În experimentele noastre folosim *patru* proteine pentru care sunt disponibile valorile RSA. Proteinele selectate sunt: **1GO1**, **1JT8**, **1L3P** și **1P1L** [BWF+00].

Setul de date pentru fiecare proteină constă din 10000 tranziții conformaționale și reprezintă evoluția structurii proteinei în intervale de timp foarte mici. În abordarea noastră, fiecare tranziție conformațională este reprezentată ca o secvență de 99 de valori numerice, reprezentând valorile RSA din structura primară a proteinei.

Înainte de aplicarea metodelor de învățare nesuppervizată, a fost utilizată o normalizare bazată pe deviație standard pentru fiecare set de date proteice.

1.1.2.1 Clustering

Pe fiecare dintre seturile de date proteice preprocesate, un experiment este efectuat pentru prima dată folosind K-means și HAC ca metode de clustering. Având 10000 conformații succesive de reprezentate de vectori RSA, le grupăm în clase folosind un hiperparametru numit **step** (s) reprezentând cardinalitatea fiecărei clase. Etapa induce un nou hiperparametru: numărul de clustere care vor fi utilizate pentru procesul de clustering (K). De exemplu, un pas $s = 2500$ ar însemna că ne așteptăm la clase $K = 4$, și anume: **Clasa 1** - conformații de la 1 la 2500; **Clasa 2** - conformații de la 2501 la 5000; **Clasa 3** - conformații de la 5001 la 7500; **Clasa 4** - conformații de la 7501 la 10000.

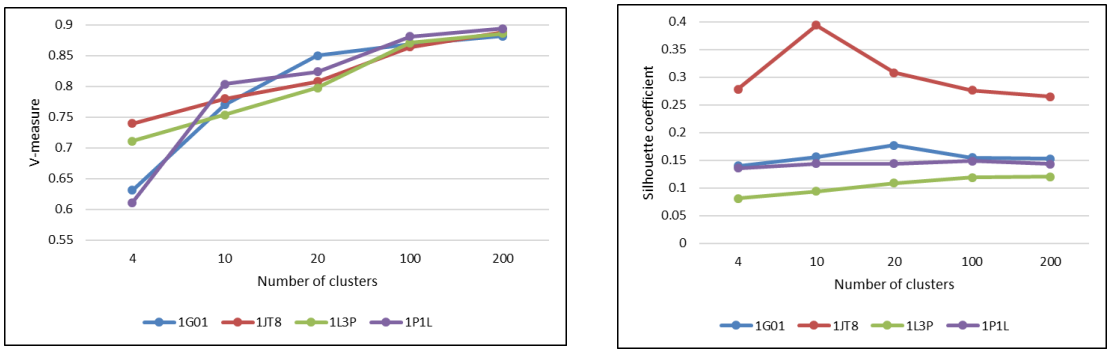
1.1.3 Rezultate și discuții

În această secțiune prezentăm rezultatele experimentelor noastre, precum și o discuție cu privire la rezultatele obținute din punct de vedere computațional și biologic. Rezultatele obținute prin experimentele de clustering vor fi prezentate, folosind metodologia experimentală descrisă în secțiunea 1.1.1. Am folosit atât clustering aglomerativ ierarhic, cât și K-means pentru procesul descris în secțiunea 1.1.2.1. Rezultatele obținute sunt prezentate în Table 1.1, unde, pentru fiecare proteină, sunt evidențiate cele mai bune valori obținute pentru *V-măsură* și *coeficientul de siluetă*. Valorile obținute pentru *V-măsură* sunt suficient de bune, variind între $0,6$ și $0,9$, ceea ce înseamnă că clasificarea se realizează relativ bine. Sunt evidențiate cele mai bune valori pentru *V-măsură* și *coeficientul de siluetă*.

Protein	**Step**	**K**	**Metodă de clustering**	*V measure*	**Coeficient de siluetă**	**Protein**	**Step**	**K**	**Metodă de clustering**	*V measure*	**Coeficient de siluetă**
1GO1	2500	4	$K-means$	0.715	0.156	1L3P	2500	4	$K-means$	0.708	0.084
			HAC	0.631	0.140				HAC	0.711	0.081
	1000	10	$K-means$	0.812	0.152		1000	10	$K-means$	0.767	0.101
			HAC	0.770	0.156				HAC	0.754	0.094
	500	20	$K-means$	0.864	**0.179**		500	20	$K-means$	0.802	**0.112**
			HAC	0.850	0.177				HAC	0.798	0.109
	100	100	$K-means$	0.869	0.160		100	100	$K-means$	0.856	0.121
			HAC	0.869	0.155				HAC	0.871	0.119
	50	200	$K-means$	0.875	0.149		50	200	$K-means$	0.880	0.117
			HAC	**0.882**	0.153				HAC	**0.886**	0.120
1JT8	2500	4	$K-means$	0.662	0.282	1P1L	2500	4	$K-means$	0.607	0.145
			HAC	0.740	0.278				HAC	0.611	0.136
	1000	10	$K-means$	0.754	0.321		1000	10	$K-means$	0.774	0.148
			HAC	0.780	0.304				HAC	0.804	0.144
	500	20	$K-means$	0.790	**0.317**		500	20	$K-means$	0.820	0.144
			HAC	0.808	0.308				HAC	0.824	0.144
	100	100	$K-means$	0.861	0.275		100	100	$K-means$	0.869	**0.151**
			HAC	0.864	0.276				HAC	0.881	0.149
	50	200	$K-means$	0.882	0.266		50	200	$K-means$	0.885	0.145
			HAC	**0.888**	0.265				HAC	**0.894**	0.143

Table 1.1: Rezultatele clustering-ului.

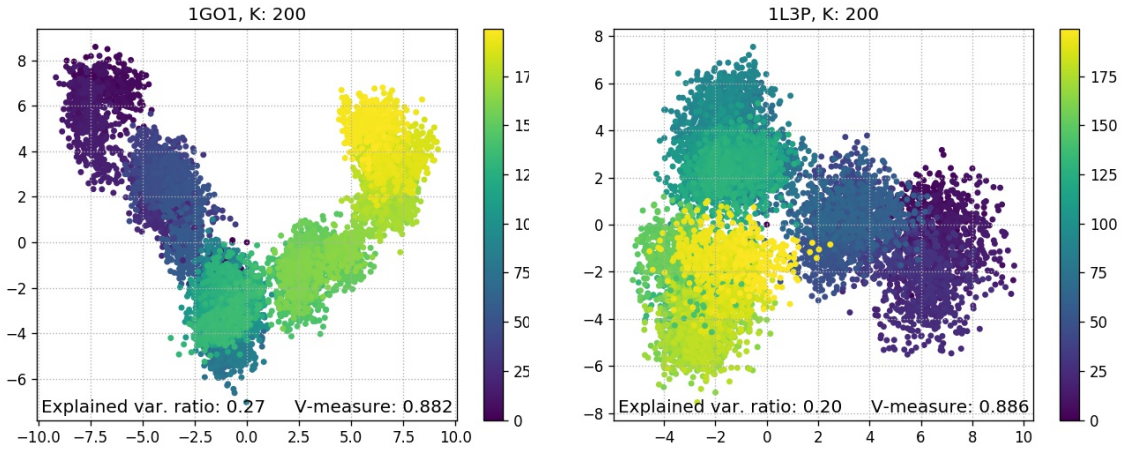
Din tabelul 1.1 observăm că, în general, grupurile furnizate de HAC sunt mai bune decât cele raportate prin metoda partițională K-means, adică *HAC* oferă coeficienți mai mari de *V-măsură* și *siluetă* pentru grupurile obținute.



(a) *V-measure.* (b) *Coeficient de siluetă.*

Figure 1.1: Rezultate pentru *HAC* având în vedere diferite numere de clusteri, pentru fiecare proteină.

Deoarece proteina suferă modificări conformaționale, anumite părți ale structurii sale sunt supuse unor modificări minore, care se reflectă în pozițiile reziduurilor de aminoacizi și, prin urmare, în valorile lor *RSA*. Astfel, conformațiile consecutive sunt destul de similare din perspectiva reprezentărilor lor considerate (valorile *RSA*).



(a) Proteina 1GO1. (b) Proteina 1L3P.

Figure 1.2: Vizualizare *HAC* a proteinelor 1GO1 și 1L3P folosing PCA.

În conformitate cu cunoștințele noastre, studiul prezentat în această teză este nou în literatura de specialitate referitoare la analiza datelor proteice conduse de *RSA*. Revista de literatură pe care am efectuat-o a relevat faptul că *învățare nesupervizată* a fost aplicată în principal pentru analiza interacțiunii dintre proteine și nu pentru analiza intra-proteină, ca în teza noastră. Cele mai multe abordări existente în literatura de specialitate referitoare la analiza intra-proteină utilizează modele *supravizate* sau *semi-supravizate* învățarea pentru clasificarea proteinelor în funcție de caracteristicile lor structurale. Doar puține abordări folosesc metode *învățare nesupervizată* pentru analiza tranzițiilor conformaționale proteice.

1.1.4 Concluzii și abordări ulterioare

În acest capitol am prezentat un studiu privind aplicarea *clusteringului* ca metodă de *clasificare nesupervizată* pentru investigarea relevanței valorilor *RSA* pentru a prezice tranzițiile interne ale proteinelor. De asemenea, analiza componentelor principale a fost explorată pentru vizualizarea datelor și folosită pentru a examina cum evoluează valorile *RSA* între tranzițiile conformaționale. Experimentele efectuate pe mai multe proteine au evidențiat faptul că valorile *RSA* se schimbă fără probleme între tranzițiile conformaționale.

Ne propunem să extindem analiza rezultatelor computaționale obținute în acest capitol din punct de vedere biologic. Pe baza studiului efectuat anterior și a investigațiilor anterioare privind analiza datelor proteice, ne propunem să avansăm cercetările noastre pentru a prezice tranzițiile conformației proteice.

1.2 Folosirea rețelelor cu autosupervizare pentru analiza dinamicii proteinelor

În secțiunea actuală, introducem diverse experimente bazate pe autoencodere (rețele cu autosupervizare) pentru analiza structurii proteinelor. Experimentele descrise în această secțiune au fost introduse în lucrarea originală [TCB+18].

Contribuția lucrărilor prezentate în acest capitol este dublă. Primul nostru obiectiv principal este de a investiga capacitatea modelelor de învățare nesupervizate. În al doilea rând, propunem două reprezentări interne pentru o proteină cu scopul de a analiza care dintre ele este mai informativă și ar conduce o rețea cu autosupervizare pentru a învăța mai bine relațiile structurale dintre proteine. Pentru a ne atinge obiectivele, experimentele

vor fi efectuate pe ansamblu de conformații pentru două proteine, *1JT8* și *1P1L*, folosind reprezentări diferite.

În concluzie, în această lucrare sunt cercetate trei întrebări de cercetare:

RQ1 Care este abilitatea *autoencoderelor* de a învăța într-un mod nesupervizat structura proteinelor și cum influențează reprezentarea internă a unei proteine în procesul de învățare?

RQ2 În ce măsură un autoencoder este capabil să păstreze asemănarea structurală dintre proteine?

RQ3 Sunt autoencoderele capabile să exprime tipare care ar fi relevante din punct de vedere biologic? Mai exact, în ce măsură rezultatele noastre de calcul sunt corelate cu perspectiva biologică?

1.2.1 Metodologie

În această secțiune prezentăm metodologia experimentală utilizată în susținerea presupunerii noastre că *autoencoderele* pot capta, din punct de vedere computațional, tipare relevante biologic în ceea ce privește modificările structurale de conformație ale proteinelor.

Pentru a răspunde la primele două întrebări de cercetare formulate anterior, experimentele vor fi efectuate în două direcții.

În primul rând, cercetăm în secțiunea 1.2.2.2 capacitatea unui *autoencoder* de a păstra structura unei proteine. Două tipuri de reprezentări vor fi luate în considerare pentru a identifica cea mai potrivită pentru analiza pe care o realizăm. Aceste reprezentări vor fi detaliate în secțiunea 1.2.1.1.

În al doilea rând, ne îndreptăm atenția către relațiile structurale ale proteinelor și analizăm dacă tehnicile de învățare auto-supervizate, cum ar fi autoencoderele, sunt capabile să detecteze anumite tipare de bază în cadrul datelor. În acest scop, realizăm un studiu de caz pe două proteine similare structural, studiu care va fi descris în continuare în secțiunea 1.2.2.3.

1.2.1.1 Reprezentarea proteinelor

O proteină este o macromoleculă cu o structură înnăscută foarte flexibilă și dinamică [MJC02] care își schimbă forma datorită atât schimbărilor externe din mediul său, cât și forțelor moleculare interne. Forma rezultată este o conformație diferită. Pentru fiecare conformație a unei proteine, în studiul nostru vor fi utilizate două reprezentări diferite ale geometriei locale a moleculei.

Prima reprezentare pentru conformația unei proteine, pe care o numim *reprezentare bazată pe unghiuri* (**Angles**), constă din stări conformaționale date de cele trei tipuri de unghiuri [PFK10].

A doua modalitate de a reprezenta o conformație proteică, denumită în cele ce urmează *reprezentare combinată* se bazează pe îmbunătățirea stărilor de conformație date de unghiuri cu valorile *RSA* ale reziduurilor de aminoacizi .

1.2.1.2 Arhitectura *autoencoderului*

În studiul curent, utilizăm autoencodere pentru a învăța reprezentări de dimensiuni inferioare semnificative pentru structurile proteinelor, având în vedere tranzițiile conformaționale ale acestora.

Vom folosi un astfel de autoencoder pentru a reduce dimensionalitatea datelor noastre. Având în vedere că unul dintre scopurile noastre este de a putea vizualiza seturile noastre de date, toate tehnicile implicate vor codifica reprezentările proteice în vectori 2 dimensionali.

țiuneMăsuri de evaluare

Pentru a determina dacă reprezentarea învățată de autoencoder păstrează conexiunile evolutive găsite în datele proteice originale, precum și pentru a identifica dacă se păstrează similaritatea structurală (dată de tranzițiile conformaționale) între proteine, am definit două măsuri de similaritate. În primul rând, măsura de asemănare intra-proteină, *IntraPS*, evaluează gradul de asemănare între conformațiile din cadrul unei proteine și vom folosi acest lucru ca o indicație a cât de bine se mențin relațiile de conformare intra-proteine în reprezentarea dimensiunii inferioare învățat de codul auto. În al doilea rând, măsura de similaritate inter-proteină, *InterPS*, evaluează gradul de corelație între două proteine, în reprezentările lor considerate și va fi utilizată pentru a identifica cât de bine sunt păstrate aceste corelații în datele rezultate, după aplicarea autoencoderului. Ambele măsuri se bazează pe măsura asemănării cosinusului, care este utilizată pentru a evalua asemănarea dintre două conformații ale unei proteine.

1.2.2 Rezultate și discuții

Experimentele pe care le-am efectuat pentru a evidenția potențialul autoencoderilor profunzi de a capta structura proteinelor vor fi prezentate în continuare, folosind metodologia experimentală prezentată în secțiunea 1.2.1.

1.2.2.1 Seturi de date

Proteinele utilizate pentru analiză sunt: 1P1L - componentă a organismelor metabolizatoare de sulf și 1JT8 - proteină implicată în traducere [BWF+00]. Acestea au fost alese pe baza disponibilității datelor (tranziții conformaționale *și RSA*), faptul că au aceeași lungime de secvență (ceea ce ne permite să efectuăm investigațiile noastre legate de RQ2) și 42 de poziții echivalente în comun (din lungimea totală de 102). [YG04].

1.2.2.2 Primul experiment

Experimentul descris mai jos este realizat cu scopul de a răspunde la prima noastră întrebare de cercetare (RQ1) cu privire la potențialul *autoencoderelor* de a învăța, fără supraveghere, structura proteinelor, precum și capacitatea lor de a descoperi tiparele biologice relevante. În același timp, investigăm dacă și modul în care reprezentarea internă a unei proteine are un impact asupra procesului de învățare.

Pentru fiecare set de date de proteine descris în secțiunea 2.2.1.1, am antrenat un număr de autoencodere (Secțiunea 2.2.1.1). Pentru autoencoder am folosit implementarea Keras disponibilă la [C+15].

Autoencoderele prezentate în secțiunea 2.2.1.1 sunt utilizate pentru a reduce dimensionalitatea datelor noastre și pentru a vizualiza seturile de date proteice. Figurile 1.3 și 1.4 înfățișează vizualizarea proteinelor din setul nostru de date utilizând autoencodere.

Datele originale transmise rețelei cu autosupervizare pentru fiecare proteină reprezintă o evoluție la timp a structurii proteinei. De la o conformație la alta, proteina poate rămâne neschimbată sau anumite părți ale acesteia pot avea modificări minore. După ce rețelele cu autosupervizare au fost antrenate, reprezentările bidimensionale ale proteinelor pe care le produc reușesc să capteze, așa cum se vede în Figurile 1.3 și 1.4, această evoluție: conformații succesive în original datele sunt înlănțuite progresiv împreună în datele de ieșire ale autoencoderului, confirmând astfel că autoencoderul învață cu exactitate tranzițiile biologice.

Mai mult, pentru a decide dacă codificatorul auto menține relațiile găsite în datele originale, folosim măsura *IntraPS*. Astfel, mai întâi calculăm aceste similitudini pentru datele originale și apoi pentru datele bidimensionale ieșite din codificare, pentru ambele reprezentări considerate. Rezultatele sunt prezentate în tabelul 1.2. Pentru fiecare proteină,

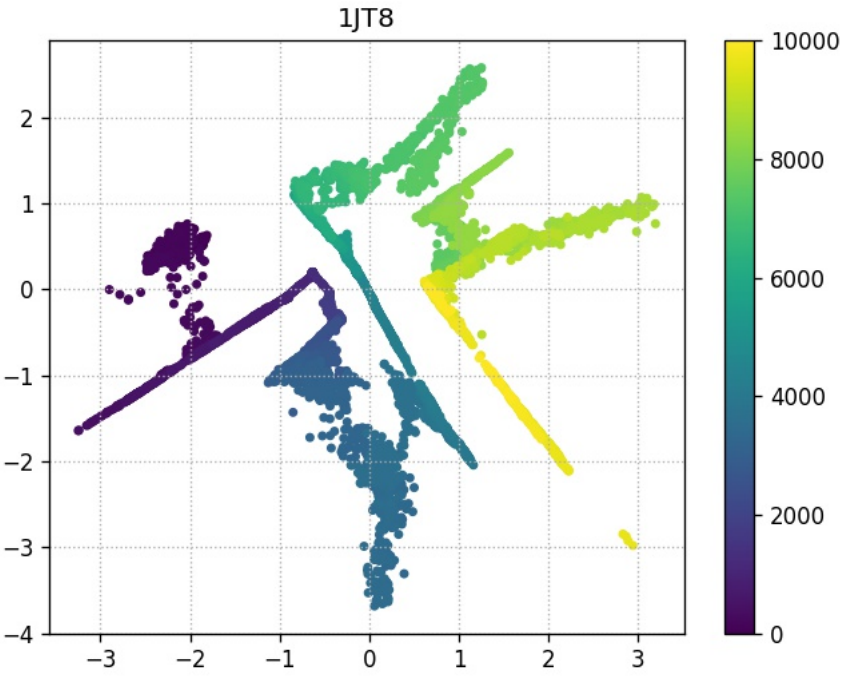


Figure 1.3: Vizualizarea proteinei 1JT8 folosind un autoencoder antrenat.

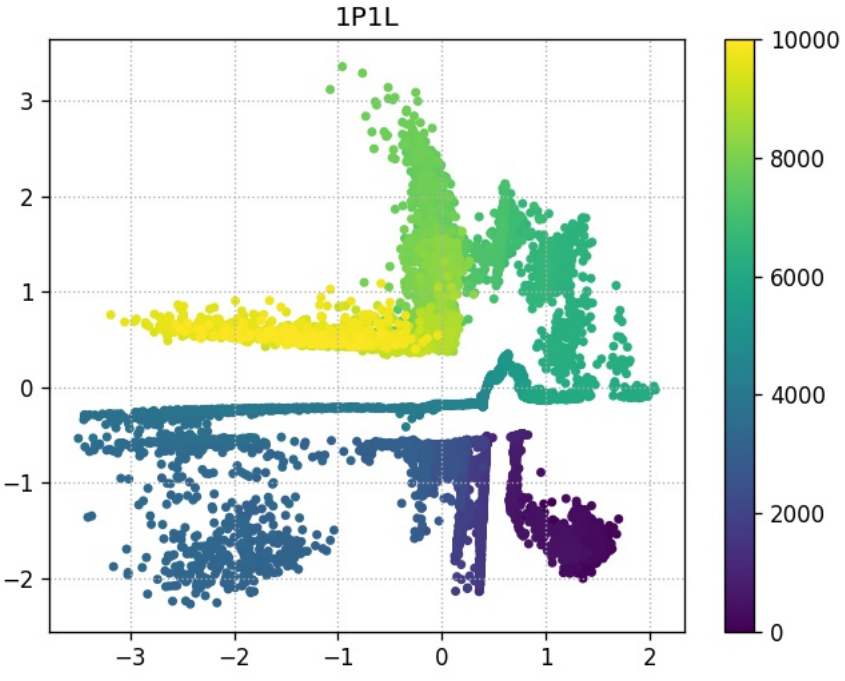


Figure 1.4: Vizualizarea proteinei 1P1L folosind un autoencoder antrenat.

pe lângă valorile pentru măsura *IntraPS*, mai prezentăm *minimum* (**Min**), *maximum* (**Max**) și *abaterea standard* (**Stdev**) a asemănărilor cosinului dintre două conformații consecutive, pentru ambele reprezentări.

Proteina		Unghiuri	Combinată	Min/Max/Stdev (COS)	
				Unghiuri	Combinată
1JT8	Original	**0.9960**	**0.9913**	0.9894/0.9995/0.0023	0.9843/0.9962/0.0022
	Codificat	**0.9939**	**0.9985**	0.9213/0.9999/0.0161	0.9573/0.9999/**0.0044**
1P1L	Original	**0.9779**	**0.9573**	0.9593/0.9896/0.0064	0.9464/0.9695/0.0054
	Codificat	**0.9912**	**0.9962**	0.9315/0.9999/0.0119	0.9661/0.9999/**0.0052**

Table 1.2: *IntraPS* pentru proteinele 1JT8 și 1P1L, folosind cele două reprezentări propuse.

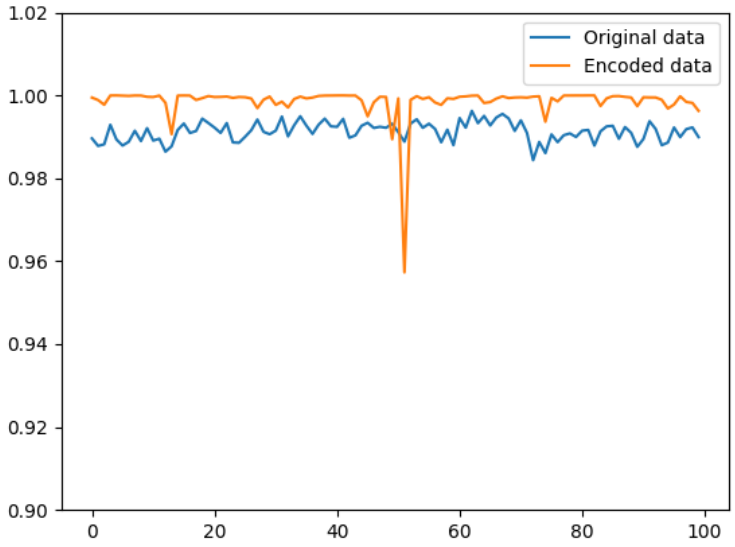


Figure 1.5: Asemănări medii comparative pentru 1JT8, pentru date originale și codificate (reprezentarea combinată).

În ceea ce privește reprezentările interne utilizate, având în vedere rezultatele obținute pentru acest experiment, concluzionăm că acestea nu influențează serios procesul de învățare. Acest lucru se poate datora reducerii semnificative a dimensionalității datelor (două dimensiuni). Cu toate acestea, pentru *reprezentare combinată*, care este mai bogată în informații decât *reprezentare bazată pe unghiuri*, s-au obținut rezultate puțin mai bune.

1.2.2.3 Al doilea experiment

Pentru a răspunde la cea de-a doua întrebare a noastră de cercetare (RQ2), menționată la începutul capitolului, s-au efectuat experimente pentru a testa dacă o rețea cu autosupervizare este capabil să învețe similitudinea structurală dintre proteine. În acest caz, ne-am concentrat pe *reprezentare combinată* pentru proteine, datorită faptului că părea să conducă la rezultate ușor mai bune (Subsecțiune 1.2.2.2).

Rezultatele sunt prezentate în tabelul 1.3.

Din tabelul 1.3 observăm că asemănările proteice calculate pe toate combinațiile de date codificate sunt mai mari decât asemănările dintre datele originale. Acest lucru sugerează că se păstrează asemănările originale existente, dar reducerea dimensionalității induce mai multe asemănări între conformațiile codificate.

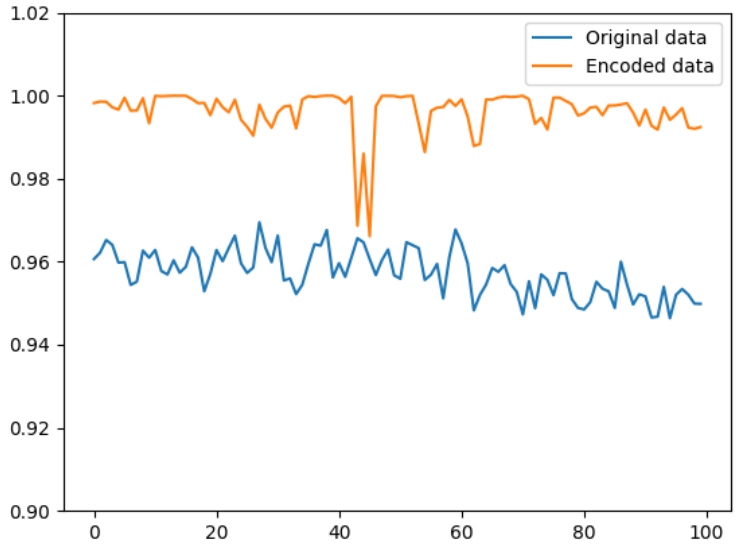


Figure 1.6: Asemănări medii comparative pentru 1P1L, pentru date originale și codificate (reprezentarea combinată).

	Reprezentare combinată
$InterPS$(1JT8, 1P1L)	**0.5438**
$InterPS$(<1JT8 codificat>, <1P1L codificat>)	**0.6367**
$InterPS$(<1JT8 codificat with 1P1L>, <1P1L codificat with 1JT8>)	**0.6478**
$InterPS$(<1JT8 codificat with 1P1L>, <1P1L codificat>)	**0.6959**
$InterPS$(<1P1L codificat with 1JT8>, <1JT8 codificat>)	**0.6297**

Table 1.3: *InterPS* pentru proteinele 1JT8 și 1P1L, folosind datele originale și codificate.

1.2.3 Concluzii și abordări ulterioare

Am realizat un studiu privind aplicarea *deep autoencoderelor* pentru o mai bună înțelegere a dinamicii proteinelor. Experimentele efectuate pe două proteine au evidențiat faptul că *autoencoderele* sunt modele eficiente nesupervizate capabile să învețe structura proteinelor, precum și să descopere asemănarea structurală dintre proteine. Mai mult, am obținut o dovadă empirică conform căreia rețelele cu autosupervizare sunt capabile să codifice tipare ascunse relevante dintr-o perspectivă biologică.

Pe baza studiului efectuat în acest capitol ne propunem să avansăm cercetările noastre pentru a prezice tranzițiile conformaților proteice folosind modele de învățare supervizate.

Capitolul 2

Modele noi de învățare automată pentru analiză inter-proteine

Proteomica este astăzi unul dintre cele mai importante și relevante domenii din biologia computațională, ridicând o mulțime de întrebări. Înțelegerea funcției și dinamicii proteinelor, precum și obținerea unor informații suplimentare în procesul de pliere a proteinelor este încă de mare interes pentru bionformatică și medicină.

În acest capitol urmărim direcția de cercetare a analizei inter proteine. Astfel, prezentăm trei contribuții principale principale care au fost publicate în lucrări originale [TCB19, TCC19, CCT19]:

- În secțiunea 2.1 examinăm utilitatea aplicării clusteringului partițional și ierarhic ca metode de clasificare nesupervizate pentru descoperirea asemănării structurale a proteinelor, pe baza informațiilor conținute în tranzițiile lor conformaționale. Cercetăm trei reprezentări pentru o proteină bazată pe distribuțiile de probabilitate ale anumitor elemente structurale în cadrul tranzițiilor conformaționale și aplicăm metode de clustering pentru a clasifica nesupervizat proteine pe baza asemănării lor structurale. Experimentele sunt efectuate pe două seturi de date proteice, iar rezultatele obținute sunt analizate și comparate cu rezultatele unor abordări similare existente. Rezultatele comparative dezvăluie faptul că, în multe cazuri, propunerea noastră este mai bună decât o lucrare anterioară în acest subiect.

- În secțiunea 2.2 investigăm problema clasificării supervizate a proteinelor în funcție de asemănarea lor structurală, pe baza informațiilor incluse în tranzițiile conformaționale ale acestora. Vă propunem abordarea *AutoSimP* constând dintr-un ansamblu de *autoencoders* pentru a prezice clasa de asemănare a unei anumite proteine, având în vedere clasa de asemănare prevăzută pentru tranzițiile conformaționale ale acesteia. Experimentele efectuate pe datele proteice reale dezvăluie eficacitatea propunerii noastre în comparație cu abordările existente similare.

- În secțiunea 2.3 introducem o nouă abordare *AnomalP* pentru detectarea tranzițiilor anomale de conformație cu proteine folosind *deep autoencoders* pentru codificarea informațiilor despre asemănarea structurală între proteinele aparținând aceleiași superfamilii. Experimentele sunt efectuate pe date proteice reale, iar rezultatele obținute subliniază potențialul auto-codificatorilor de a învăța tipare relevante biologice, cum ar fi caracteristicile structurale ale proteinelor și că sunt utile pentru detectarea conformațiilor sau proteinelor care pot fi anomale în raport cu o superfamilie. Studiul efectuat în această secțiune are ca scop să ofere o perspectivă mai bună a similarității structurale a proteinelor, cu scopul mai larg de a învăța să prezice tranzițiile conformaționale ale proteinelor.

Secțiunea 2.1 este structurată după cum urmează. Vă propunem trei tehnici diferite de reprezentare a proteinelor în secțiunea 2.1.1. Având scopul de a compara eficiența lor, efectuăm experimente folosind fiecare dintre ele și raportăm performanțele obținute în secțiunea 2.1.3.

Secțiunea 2.2 reprezintă o extensie a metodei anterioare și este structurată după cum urmează. Metoda care se bazează pe un ansamblu de codificatoare automate profunde este prezentată în secțiunea 2.2.1. Apoi, efectuăm experimente și raportăm rezultatele obținute în secțiunea 2.2.2.

Secțiunea 2.3 prezintă o metodă de detectare a anomaliilor bazată pe codificatoare auto adânci și este structurată după cum urmează. *AnomalP*, abordarea pentru detectarea tranzițiilor conformaționale anomale este descrisă în secțiunea 2.3.1.

2.1 Modele de clustering pentru descoperirea similarității structurale a proteinelor

În această secțiune, cercetăm utilitatea aplicării clusteringului partițional și ierarhic ca metode de clasificare nesupervizate pentru descoperirea similarității structurale a proteinelor, pe baza informațiilor conținute în tranzițiile conformaționale ale acestora. Experimentele descrise în această secțiune au fost introduse în lucrarea originală [TCB19].

Contribuția secțiunii este rezumată în cele ce urmează. Principalul obiectiv este de a sublinia eficiența metodei de clustering partițional și ierarhic pentru detectarea, bazată pe tranzițiile conformaționale ale proteinelor, relațiilor structurale dintre acestea.

În această lucrare răspundem la următoarele întrebări de cercetare:

RQ1 Care este eficacitatea utilizării metodelor de clustering pentru a clasifica proteinele în funcție de relațiile structurale dintre ele?

RQ2 Cum se compară abordarea bazată pe clustering introdusă în această lucrare cu lucrările existente legate de detectarea asemănării structurale inter-proteine?

2.1.1 Metodologie

Prezentăm în continuare metodologia pe care se bazează studiul nostru. Secțiunea 2.1.2 prezintă modelele vectoriale pe care le propunem pentru reprezentarea unei proteine folosind distribuția probabilității elementelor alfabetului structural în tranzițiile conformaționale ale proteinei.

2.1.2 Reprezentări ale proteinelor

Avem în vedere în următoarele trei reprezentări vectoriale pentru proteine bazate pe distribuțiile literelor SA în tranzițiile lor conformaționale.

Să considerăm că o proteină $Prot$ de lungime n este vizualizată ca o secvență de caractere peste alfabetul $\mathcal{A} = \{G, P, A, V, L, I, M, C, F, Y, W, H, K, R, Q, N, E, D, S, T\}$ din 20 de litere reprezentând aminoacizi: $Prot = a_1 a_2 \dots a_n$, unde $a_i \in \mathcal{A}, \forall i \in \{1, 2, \dots, n\}$ [BPC+17]. Pentru o proteină, sunt date mii de conformații diferite (reprezentate așa cum este descris mai sus) obținute prin simulări de dinamică moleculară. O conformație a proteinei poate fi transformată în reprezentarea SA, alfabetul structural SA fiind compus din cele 25 de litere:

$$SA = \{A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, V, W, X, Y\}$$

. Să notăm cu l_i $1 \leq i \leq 25$ literele din alfabetul SA.

În consecință, pentru o proteină $Prot$ este dat un număr mare de m de conformații determinate experimental. Astfel, proteina $Prot$ se caracterizează printr-o secvență de conformații m,

$$Seq_{Prot} = (cf_1^{Prot}, cf_2^{Prot}, \ldots, cf_m^{Prot}),$$

unde

$$cf_i^{Prot} = (cf_{i1}^{Prot}\ cf_{i2}^{Prot} \ldots cf_{i,n-3}^{Prot}),\ cf_{ik}^{Prot} \in SA,\ 1 \leq i \leq m$$

.

Prima repezentare (R1)

Prima reprezentare pentru o proteină $Prot$ este un vector de frecvență pe care l-am introdus anterior în [BPC+17] și este construit după cum urmează. Pentru fiecare din cele 25 de litere din alfabetul structural, se calculează probabilitatea pr_{l_i} de apariție a literei l_i în tranzițiile conformaționale ale proteinei $Prot$. Astfel, $Prot$ este vizualizat ca un vector dimensional de 25 care conține probabilitatea apariției simbolurilor SA în proteina dată, $Prot = (pr_{l_1}, pr_{l_2}, \ldots, pr_{l_{25}})$.

A doua reprezentare (R2)

Având în vedere conformațiile date pentru o proteină $Prot$, se poate calcula un vector de distribuție, care stochează informații despre distribuția literelor SA în conformațiile proteinei. Vă propunem următorul calcul pentru vectorul de distribuție. Pentru fiecare conformație cf_i^{Prot} a proteinei $Prot$ și pentru fiecare dintre literele de 25 din alfabetul structural $lt \in SA$, calculăm probabilitatea de p_{lt}^i de apariție de literă lt în tranziția conformațională cf_i^{Prot}. Astfel, proteina $Prot$ poate fi vizualizată ca un vector dimensional de 25 $\cdot m$ care conține probabilitățile apariției simbolurilor SA în tranzițiile sale conformaționale, $Prot = (p_{l_1}^1, \ldots, p_{l_{25}}^1, p_{l_1}^2, \ldots, p_{l_{25}}^2, \ldots p_{l_1}^m, \ldots, P_{l_{25}}^m)$.

A treia reprezentare (R3)

A treia reprezentare pe care o propunem pentru o proteină $Prot$ are în vedere nu numai probabilitatea literelor SA în tranzițiile conformaționale de $Prot$, ci și frecvența tuturor bigramelor și trigramelor posibile compuse de literele din alfabetul structural.

Având în vedere o anumită literă SA l (de ex. litera A), există 651 posibile uni, bi și trigrame începând cu litera l (de ex. $A, AA, AB, \ldots AZ, AAA, AAB, \ldots AZZ$). În consecință, se pot forma un număr de 16275 uni, bi și trigrame, luând în considerare toate cele 25 de litere SA. Să notăm prin $Seq = (seq_1, \ldots, seq_{16275})$ secvența tuturor posibilelor n -gram $(1 \leq n \leq 3)$.

O proteină $Prot$ poate fi apoi vizualizată ca un vector numeric dimensional de 16275 $Prot = (prot_1, prot_2, \ldots, prot_{16275})$, unde $prot_i$ reprezintă probabilitatea apariției de n -gram seq_i în toate tranzițiile conformaționale ale proteinei.

2.1.2.1 Modelele de clustering

Având în vedere modelele de reprezentare propuse anterior pentru o proteină, presupunem că avem un set de proteine $SP = \{Prot_1, Prot_2, \ldots, Prot_r\}$, fiecare proteină $Prot_i$ fiind reprezentată ca vector numeric multidimensional, așa cum este descris de mai sus. Cu scopul principal de a grupa proteinele din SP, astfel încât un grup conține proteine care sunt similare din punct de vedere structural, sunt aplicați doi algoritmi de clustering: $k-means$ și HAC. Funcția de distanță folosită pentru a măsura disimilaritatea dintre proteine este *distanța euclidiană* între vectorii lor de înaltă dimensiune. Au fost luate în considerare diverse funcții de distanță, dar, în general, distanța euclidiană a funcționat mai bine, prin urmare, ne vom concentra asupra acestei funcții.

2.1.3 Rezultate experimentale

Două seturi de date proteice vor fi utilizate în continuare în experimentele noastre, aplicând metodologia introdusă în secțiunea 2.1.1. Ambele seturi de date au fost utilizate într-un studiu anterior [ATC18] privind detectarea similarității inter-proteine. Fiecare proteină din seturile de date se caracterizează printr-o secvență de 10000 conformații determinate experimental (adică $m = 10000$). Proteinele au lungimi diferite, variind de la 99 la 668 aminoacizi.

2.1.3.1 Seturi de date

Dataset D1 este format din *seven* proteine (coduri: 1ASH, 1DLW, 1ECA, 1C52, 1CCR, 1APQ, 1COU în *PDB* [BWF+00]), prelevate din trei superfamilii diferite (1.10.490.10, 1.10. 760.10, 2.10.25.10).

Datetul D2 se extinde pe D1 și este format din 58 de proteine aparținând la nouă familii diferite.

2.1.3.2 Rezultate

Pentru a răspunde la întrebarea de cercetare RQ1 și pentru a testa potențialul de K - *means* și HAC pentru a clasifica proteinele în funcție de similitudinea lor structurală, modelele de clustering introduse în secțiunea 2.1.2.1 sunt aplicate pe seturi de date D1 și D2 descrise mai sus. Ambele reprezentări proteice introduse în secțiunea 2.1.2 sunt utilizate în experimente. Partițiile rezultate sunt evaluate folosind măsurile prezentate în secțiunea 1.2.1.2.

Rezultatele obținute sunt prezentate în tabelul 2.1, unde, pentru fiecare set de date, sunt evidențiate cele mai bune valori obținute pentru *V-measure* și *coeficientul de siluetă*. Observăm valori relativ bune pentru ambele măsuri de evaluare.

Set de date	**Reprezentare**	**Metoda de clustering**	***V-measure***	***Coeficient de siluetă***
D1	R1	*K-means*	1	0.60 ± 0.00
		HAC	1	0.60
	R2	*K-means*	1	0.51±0.00
		HAC	1	0.51
	R3	*K-means*	1	0.59±0.00
		HAC	1	0.59
D2	R1	*K-means*	0.68 ± 0.01	0.28 ± 0.01
		HAC	0.70	0.29
	R2	*K-means*	0.64±0.02	0.21 ± 0.01
		HAC	0.67	0.22
	R3	*K-means*	0.66±0.01	0.23±0.01
		HAC	0.62	0.22

Table 2.1: Rezultatele clusteringului.

Din tabelul 2.1 observăm că pentru cel de-al doilea set de date, grupurile furnizate de HAC sunt mai bune decât cele raportate de metoda partițională K - *means*, obținând mai mare V - *measure* și *coeficienți de siluetă*. În ceea ce privește primul set de date, ambii algoritmi conduc la rezultate similare pentru măsurile de evaluare.

În general, considerăm HAC-ul ca fiind cel mai performant algoritm. Tabelul 2.1 dezvăluie că a obținut cea mai bună performanță pentru ambele seturi de date. Reprezentarea care oferă cele mai bune scoruri de evaluare este cea bazată pe distribuția literelor SA

în toate tranzițiile conformaționale (R1). Acesta este un rezultat interesant, deoarece reprezentările R2 și R3 păreau a fi mai precise, codificând mai multe informații despre o proteină decât R1. O posibilă explicație pentru acest rezultat poate fi dimensionalitatea ridicată a reprezentărilor vectoriale corespunzătoare R2 și R3 care pot afecta negativ performanțele algoritmilor de clustering.

2.1.4 Concluzii

Am prezintat un studiu privind aplicarea clustering-ului partițional și ierarhic pentru clasificarea neobservată a proteinelor în funcție de asemănarea lor structurală. Scopul nostru principal a fost de a găsi dovezi empirice pentru două întrebări de cercetare legate de analiza datelor inter-proteine. În primul rând, luând în considerare trei reprezentări vectoriale pentru o proteină bazată pe tranzițiile conformaționale ale acesteia codificate folosind un alfabet structural [PFK10], a fost investigată utilitatea grupării. În al doilea rând, am comparat abordările noastre de clustering cu activități similare.

Lucrările viitoare vor viza investigarea reprezentărilor vectoriale alternative pentru o proteină bazată pe tranzițiile conformaționale ale acesteia. În plus, intenționăm să extindem evaluarea experimentală pe alte seturi de date proteice mai mari pentru o mai bună validare a concluziilor studiului nostru.

2.2 *AutoSimP*: O metodă de predicție a similarității structurii proteinelor folosind ansamble de autoencodere

Această secțiune investighează problema clasificării supervizate a proteinelor în funcție de asemănarea lor structurală, pe baza informațiilor incluse în tranzițiile conformaționale ale acestora. Experimentele au fost introduse în lucrarea originală [TCC19].

Scopul nostru principal este de a introduce o abordare de învățare supervizată *AutoSimP* bazată pe un ansamblu de codificatoare auto pentru a prezice superfamilia din care face parte o proteină. Predicția se face pe baza asemănării dintre conformațiile proteinei și conformațiile proteinelor din fiecare superfamilie (codată într-un *autoencoder*).

2.2.1 Metodologie

Prin intermediul *autoencoders*, ne propunem să testăm capacitatea lor de a păstra structura proteinelor și de a analiza dacă acestea sunt capabile să detecteze anumite relații structurale subiacente din datele proteice. Experimentele vor fi proiectate pentru a testa în ce măsură datele proteice de dimensiuni inferioare codificate sunt în conformitate cu veridicitatea biologică și pentru a stabili dacă auto-codificatoare sunt eficiente în învățarea caracteristicilor structurale ale proteinelor.

AutoSimP implică patru pași ilustrați în figura 2.2: *reprezentare și preprocesare a datelor*, *antrenare* și *testare* (*evaluare*). Etapele principale ale *AutoSimP* vor fi detaliate. Ansamblul de modele reprezintă un set de rețele neuronale convoluționale, care iau ca intrare un vector întreg cu număr real reprezentând o conformație.

2.2.1.1 Reprezentarea datelor și preprocesare

Să considerăm că $\mathcal{F} = \{F_1 \dots F_{nf}\}$ sunt superfamilii proteice. O superfamilie F_i constă din proteine n_i, adică $F_i = \{p_1^i \dots p_{n_i}^i\}$. Pentru o proteină p_j^i ($\forall 1 \leq j \leq n_i$) se dau un număr de m de diferite conformații obținute prin simulări de dinamică moleculară.

Pentru fiecare superfamilie F_i ($1 \leq i \leq nf$), un set de date D_i este construit folosind toate conformațiile pentru toate proteinele din acea superfamilie. Astfel, setul de date D_i

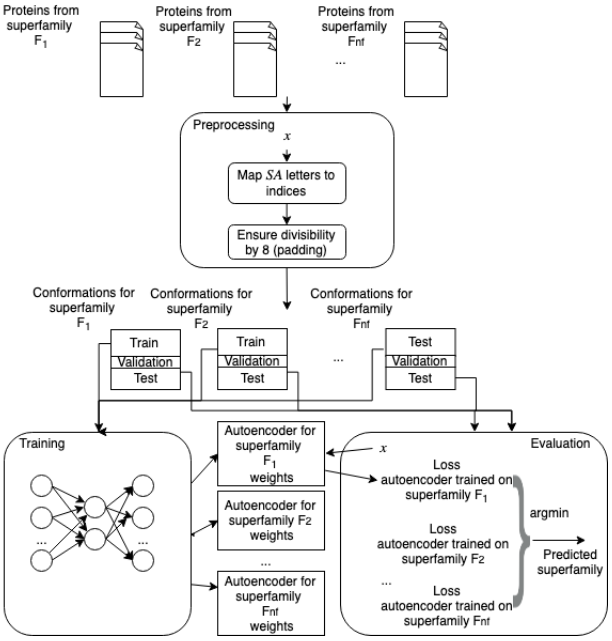


Figure 2.1: Metoda *AutoSimP*.

constă din $m \cdot n_i$ conformații, adică m conformații pentru fiecare proteină p_j^i $(1 \leq j \leq n_i)$ din i - superfamilia.

Arhitectura autoencoderului

În studiul curent, utilizăm autoencoders pentru a învăța reprezentări semnificative pentru structurile proteinelor din tranzițiile lor conformaționale.

Vom folosi un astfel de model A_i pentru a învăța o reprezentare dimensională inferioară pentru proteine din superfamilia F_i. Principalele scopuri este de a învăța reprezentări semnificative specifice superfamiliilor.

2.2.1.2 Testare

După ce a fost instruit ansamblul de modele, *AutoSimP* este evaluat folosind 24 % din fiecare set de date D_i $(\forall 1 \leq i \leq nf)$, adică 24% conformații pentru fiecare proteină din fiecare superfamilie F_j care nu au fost folosite în timpul antrenamentului.

Când testarea se efectuează la nivel de conformație, o conformație c este clasificată de *AutoSimP* ca aparținând superfamiliei F_i astfel încât $i = _{j=1,nf}\ L_j(\widehat{c}, c)$. În formula precedentă, am notat cu $L_j(\widehat{c}, c)$ valoarea erorii calculate pentru conformația c de către codificatorul auto A_j corespunzător superfamiliei j.

Când testarea este efectuată la nivel de proteină, o proteină p reprezentată în testare a fost setată ca o secvență $(c_1, c_2, \ldots, c_t)$ de conformații ($t = \frac{24 \cdot m}{100}$ după cum s-a menționat anterior) va fi clasificat ca aparținând superfamiliei F_i al cărui autoencoder A_i minimizează pierderea medie a conformațiilor sale de t, adică $i = _{j=1,nf}\ \frac{\sum_{k=1}^{t} L_j(\widehat{c_k}, c_k)}{t}$.

2.2.2 Rezultate

Cu scopul de a răspunde la întrebările de cercetare vor fi efectuate experimente pe nouă superfamilii proteice, folosind metodologia introdusă în secțiunea 2.2.1.

Abordarea *AutoSimP* a fost aplicată pe datele proteice descrise în secțiunea 2.3.2.1 urmând metodologia introdusă în secțiunea 2.2.1. Pentru fiecare superfamilie F_i ($1 \leq i \leq 9$), un set de date D_i este format din toate conformațiile pentru toate proteinele pentru superfamilie. Pentru fiecare proteină din setul de date D_i, se utilizează 6000 conformații pentru antrenament, 1600 pentru validare și restul de 2400 pentru testare. Pentru evidențierea generalității *AutoSimP*, am crescut nivelul de granularitate aplicându-l și la un nivel de conformație, adică pentru predicţia superfamiliei pentru o anumită conformație proteică. Experimentele au fost efectuate pe același set de date (Secțiunea 2.3.2.1), urmând aceeași metodologie ca și pentru experimentele efectuate la nivel de proteine.

Nivel	**Metrică**	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	F_9
Protein	*Precizie*	1	1	1	0.857	0.6	1	0.857	0.444	1
	Recall	1	1	0.8	1	0.5	1	0.666	0.666	1
	F-measure	1	1	0.888	0.923	0.545	1	0.75	0.533	1
Conformație	*Precizie*	0.997	0.969	0.994	0.809	0.695	0.983	0.858	0.433	0.996
	Recall	0.963	1	0.803	0.955	0.508	0.999	0.638	0.696	1
	F-measure	0.980	0.984	0.888	0.876	0.587	0.991	0.732	0.534	0.998

Table 2.2: Rezultate experimentale.

Rezultatele din tabelul 2.2 dezvăluie valori ridicate ale $F - measure$ pentru toate superfamiliile, cu excepția lui F_5 și F_8. Performanțele mai scăzute ale acestor superfamilii se pot datora faptului că conformațiile pentru proteinele lor sunt foarte similare cu conformațiile din proteinele aparținând altor superfamilii. Astfel, aceste conformații sunt greu de diferențiat de modele. Anumite investigații vor fi efectuate în această direcție.

2.2.3 Concluzii

Am introdus în această lucrare o abordare de învățare supervizată *AutoSimP* constând dintr-un ansamblu de *deep autoencoders* pentru clasificarea proteinelor în superfamilii pe baza informațiilor incluse în tranzițiile lor conformaționale. Experimentele efectuate pe 57 de proteine aparținând a 9 superfamilii au evidențiat eficacitatea auto-codificatorilor și capacitatea lor de a descoperi structura proteinelor și asemănarea lor structurală. Experimentele au fost efectuate atât la nivel de proteine, cât și la nivel de conformație, subliniind astfel generalitatea propunerii noastre.

2.3 *AnomalP*: O metodă de detecție a conformațiilor proteice anomale folosind deep autoencoders

În această secțiune, introducem o nouă abordare *AnomalP* pentru detectarea tranzițiilor conformaționale proteice anomale folosind *deep autoencoders* pentru codificarea informațiilor despre asemănarea structurală între proteinele aparținând aceleiași superfamilii. Studiul efectuat în această secțiune are ca scop să ofere o perspectivă mai bună a similarității structurale a proteinelor, cu scopul mai larg de a învăța să prezice tranzițiile conformaționale ale proteinelor. Experimentele au fost introduse în lucrarea originală [CCT19].

Principala contribuție a studiului constă în investigarea utilizării codificatoarelor auto pentru a decide dacă o anumită conformație a unei proteine este diferită structural față de superfamilia sa, fiind astfel posibil să reprezinte o anomalie în ceea ce privește superfamilia respectivă. Tranzițiile conformaționale ale proteinelor sunt reprezentate în studiul nostru folosind litere din *alfabet structural* (SA) [PFK10]. Întrebările de cercetare care reprezintă accentul activității noastre sunt următoarele:

RQ1 În ce măsură pot fi utilizate autoencoderele pentru detectarea conformațiilor proteice care sunt probabil anomale în raport cu superfamilia proteinei, adică conformațiile a căror structură nu seamănă cu informațiile codificate ale acestora?

RQ2 Cum influențează reprezentarea conformațiilor proteinelor asupra performanței predictive a procesului de detectare?

RQ3 Cum se aplică abordarea introdusă pentru a răspunde la întrebările anterioare de cercetare la un nivel proteic, adică pentru a decide dacă o proteină poate aparține unei anumite superfamilii, luând în considerare disimilarea dintre tranzițiile conformaționale și cele codificate informații structurale despre superfamilie?

Pentru a răspunde la prima întrebare de cercetare, introducem o abordare de învățare supervizată *AnomalP* folosită pentru a oferi probabilitatea ca o conformație să fie anomală în ceea ce privește superfamilia de proteine. Predicția se face luând în considerare gradul de disimilitate al conformației cu privire la toate conformațiile proteinelor din superfamilia dată, astfel cum este codat într-un *autoencoder*.

A doua întrebare de cercetare va fi investigată luând în considerare o reprezentare vectorială alternativă pentru conformații, înlocuind literele SA cu unghiurile dintre aminoacizii consecutivi subiacenți din structura primară a proteinei (adică unghiul de torsiune al tuturor celor patru atomi [PFK10] împreună) cu atomii de carbon alfa ai aminoacizilor).

Pentru a evidenția generalitatea lui *AnomalP* și pentru a răspunde la a treia întrebare de cercetare, *AnomalP* va fi aplicat la un nivel de granularitate mai mare, pentru a decide dacă o anumită proteină aparține sau nu unei superfamilii date.

2.3.1 Metodologie

În continuare, introducem o abordare *AnomalP* pentru detectarea tranzițiilor anomale proteice conformaționale utilizând modele profunde pentru codificarea informațiilor despre asemănarea structurală între proteinele aparținând aceleiași superfamilii.

2.3.1.1 Model teoretic

Problema pe care ne concentrăm este o problemă de clasificare binară. Având în vedere un set $\mathcal{F}$ de superfamilii de proteine, $\mathcal{F} = \{F_1 \dots F_{nf}\}$, ne propunem să prezicem dacă o anumită conformare proteică aparține sau nu unei anumite superfamilii F_i, și anume pentru a detecta probabilitatea de a fi anomal în ceea ce privește F_i. Pentru a decide dacă o conformație poate reprezenta o anomalie în raport cu F_i, calculăm gradul de disimilitate al conformației date cu privire la toate conformațiile proteinelor de la F_i, astfel cum sunt codate într-un *autoencoder*.

Pe lângă o granularitate la nivel conformațional, *AnomalP* poate fi, de asemenea, utilizat la un nivel de proteină pentru a prezice dacă o proteină (caracterizată prin tendințele conformaționale ale acesteia) aparține sau nu unei superfamilii date.

AnomalP are scopul de a demonstra empiric că autoencoderele sunt capabile să învețe asemănarea structurală și relațiile dintre proteine, precum și să fie utilizate ca un detector de anomalii, adică să determine dacă o conformație / proteină este probabil să fie anormală cu respectarea unei anumite superfamilii. Există trei etape principale ale *AutoSimP* așa cum este ilustrat în figura 2.2:

1. **Reprezentarea datelor și preprocesarea**.
2. **Anrenare**.
3. **Testare**.

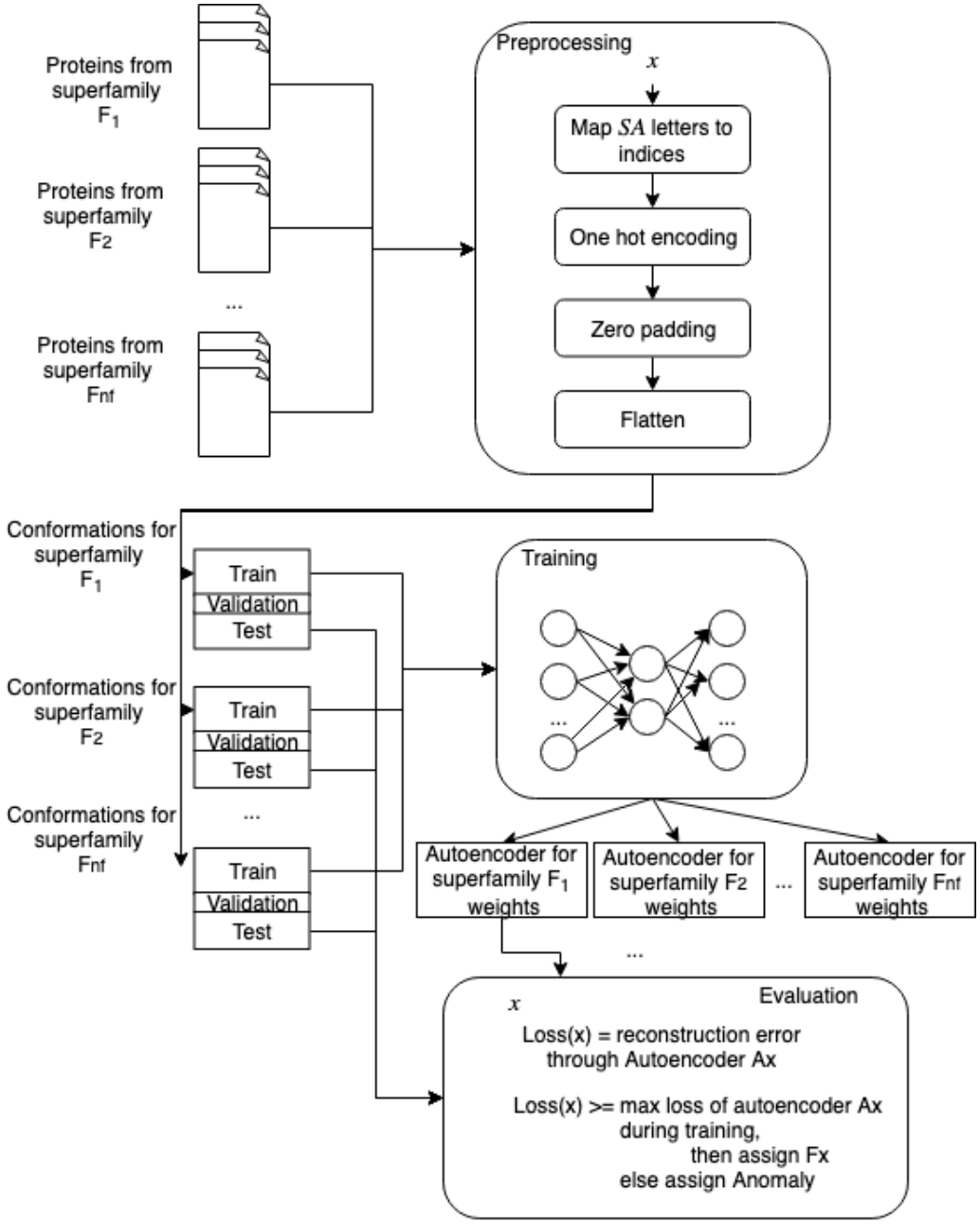


Figure 2.2: Abordarea *AnomalP*.

Arhitecturea autoencoderului

Autoencoderele folosite pentru învățarea tiparelor pentru fiecare superfamilie specifică de proteine sunt formate dintr-un codificator și un decodor. Stratul ascuns are o dimensiune de 10 în cazul în care este utilizată reprezentarea bazată pe rang, respectiv 25 atunci când este utilizată reprezentarea bazată pe unghiuri. Am experimentat și cu alte dimensiuni intermediare și am încercat să reducem dimensionalitatea până la punctul în care performanța nu a fost satisfăcătoare.

Funcția eroare care este utilizată pentru antrenarea codificatoarelor auto este *binary crossentropy*. Pentru reprezentarea bazată pe rang, această funcție este utilizată și la calcularea probabilității de anomalie. Cu toate acestea, pentru reprezentarea bazată pe unghiuri, eroarea medie absolută este monitorizată împreună cu *binary crossentropy* în timpul antrenamentului, iar prima este utilizată la calcularea probabilităților.

2.3.1.2 Clasificare folosind *AnomalP*

După ce modelul de clasificare *AnomalP* a fost construit, în stadiul de testare, vom decide dacă o anumită conformație c este probabil să fie anomală în ceea ce privește superfamilia F_i dacă eroarea autoencoderului A_i calculat pentru conformația respectivă este mai mare decât eroarea maximă obținută în timpul antrenamentului de A_i. Intuitiv, acest lucru înseamnă că conformația c este probabil diferită de informațiile structurale codificate în A_i și caracterizând superfamilia F_i.

2.3.2 Evaluare experimentală

Cu scopul de a răspunde la întrebările de cercetare RQ1 și RQ2 este furnizată în continuare o evaluare experimentală a clasificatorului nostru *AnomalP* pentru detectarea tranzițiilor conformaționale anomale. Performanța *AnomalP* este experimentată pe nouă superfamilii proteice.

2.3.2.1 Set de date

Setul de date utilizat în experimentele noastre constă din 57 de proteine aparținând nouă superfamilii și a fost folosit anterior în literatură pentru analiza intra și inter-proteină [ATC18, TCB19]. Setul de date a fost obținut din baza de date MoDEL disponibilă la [MDH+10].

2.3.2.2 Rezultate

Modelele *AnomalP* au fost aplicate pe datele proteice descrise în secțiunea 2.3.2.1 urmând metodologia introdusă în secțiunea 2.3.1.

2.3.3 Discuție

În secțiunea 2.3.3.1 este prezentată o analiză a rezultatelor experimentale din secțiunea 2.3.2.2 obținută prin aplicarea clasificatorului AnomalP pentru detectarea tranzițiilor anomale.

2.3.3.1 Analiza abordării *AnomalP*

Evaluarea experimentală a propunerii noastre de *AnomalP* realizată pe proteine reale (Secțiunea 2.3.2.2) a evidențiat performanțele sale foarte bune în detectarea probabilității unei tranziții conformaționale a unei proteine pentru a fi o anomalie în ceea ce privește superfamilia proteinei. Presupunerea că am utilizat în evaluarea noastră actuală că superfamilia proteinei este cunoscută atunci când aplicăm *AnomalP* nu este o limitare a propunerii noastre. Într-un scenariu real, în care ni se oferă pur și simplu o tranziție conformațională a unei proteine, fără să știm superfamilia acesteia, putem aplica în prealabil o abordare pentru determinarea superfamiliei proteinei. Am introdus anterior o astfel de abordare, numită *AutoSimP* [TCC19], care folosește un ansamblu de autoencodere pentru codificarea informațiilor structurale despre superfamiliile proteinelor cu scopul de a prezice superfamilia unei proteine noi.

Din Figura 2.3 și Tabelul 2.2 observăm o performanță mai slabă de *AnomalP* folosind reprezentarea bazată pe *unghiuri*, atât în ceea ce privește *precizia* cât și AUC, pentru superfamilile F_2 și F_8. Performanța mai scăzută a acestor superfamilii se poate datora faptului că informațiile codificate despre conformațiile pentru proteinele lor sunt foarte similare cu conformațiile provenite din proteine aparținând altor superfamilii. Astfel, aceste conformații sunt greu de diferențiat de codificatoarele auto. Anumite investigații vor fi efectuate în această direcție.

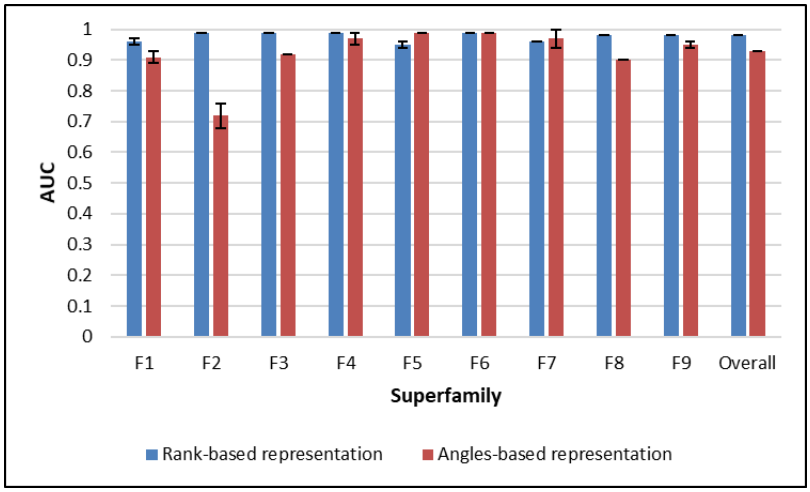


Figure 2.3: Valorile AUC obținute de *AnomalP* folosind reprezentările pentru conformații bazate pe rang respectiv unghi. Barele de eroare reprezintă intervalul de încredere 95%.

2.3.4 Concluzii

În această secțiune am propus o abordare bazată pe învățare automată *AnomalP* care a folosit autoencodere pentru a determina dacă o anumită conformație proteică este diferită structural de superfamilia proteinei, fiind astfel probabil să reprezinte o anomalie. Generalitatea *AnomalP* a fost evidențiată prin aplicabilitatea sa la nivel proteic, nu numai la nivelul conformației. Evaluarea experimentală efectuată pe proteine reale aparținând la nouă superfamilii a relevat o performanță predictivă foarte bună pentru abordarea propusă. Astfel, am obținut dovezi empirice conform cărora autoencoderele sunt capabile să codifice cu exactitate relațiile dintre proteinele incluse în aceeași superfamilie.

Capitolul 3

Modele de învățare profundă pentru vedere artificială

Clasificarea imaginilor este un exemplu bine cunoscut al unei probleme complexe care poate fi abordată folosind învățarea profundă. În ultima perioadă, rețelele neuronale convoluționale au fost aplicate cu succes pentru a rezolva problema. Astfel, dezvoltarea recentă a abordărilor de învățare profundă a îmbunătățit performanța sistemelor de recunoaștere vizuală.

Din perspectiva viziunii computerizate, dezvoltarea unui clasificator de imagini constă în scrierea unui algoritm care este capabil să clasifice imaginile în clase distincte. Pentru îmbunătățirea robusteței acestor clasificatori, cercetătorii au propus o abordare bazată pe date. Așadar, dintr-o perspectivă *ML*, în loc să descriem în mod explicit cum arată fiecare categorie de imagini, oferim algoritmului mostre etichetate pentru fiecare clasă de imagini. Algoritmul de învățare își va regla parametrii pentru a învăța aspectul vizual al fiecărei clase de imagine.

În acest capitol evidențiem trei contribuții principale principale care au fost publicate în lucrări originale [Tel17, TD19, DT19]:

- În secțiunea 3.1 investigăm posibilitatea de a utiliza unele tehnici de învățare supervizată pentru a construi modele capabile să efectueze cu exactitate analiza scrisă a semnăturilor. Rezultatele raportate în faza de testare a modelului obținut sunt încurajatoare pentru lucrări suplimentare. Decizia dacă o semnătură scrisă de mână este legitimă sau că a fost falsificată este o sarcină foarte complexă. Mai multe metode au fost încercate de experții în grafologie pentru a detecta o astfel de fraudă. Cu toate acestea, este evident că este foarte greu de efectuat o astfel de clasificare.

- În secțiunea 3.2 vă propunem o soluție ușoară pentru rezolvarea unei probleme de clasificare a imaginii, și anume recunoașterea fructelor. Soluția este testată pe un set de date deschis. Obținem performanțe de ultimă generație atât în ceea ce privește acuratețea clasificării, cât și viteza de execuție. Observăm că recent, direcțiile de cercetare s-au concentrat mai mult pe dezvoltarea de modele ușoare care încă pot atinge performanțe bune de clasificare.

- Pe lângă clasificarea imaginilor, există o altă direcție importantă de cercetare, și anume localizarea obiectelor. Astfel, în secțiunea 3.3 prezentăm o nouă metodă de preprocesare bazată pe învățare profundă, pentru a detecta și desprinde documente în imagini digitale. Lucrarea noastră intenționează să îmbunătățească performanțele de recunoaștere optică, în special pe cadre care sunt înclinate (ușor rotite) sau care au fundaluri înghesuite. Metoda propusă obține rezultate bune de detectare și deșurare a documentelor pe un set de date cu fotografii ale încasărilor în numerar.

- În secțiunea 3.4 introducem o nouă abordare pentru gruparea proteinelor bazată pe asemănarea structurii interne a acestora. Vă propunem un sistem bazat pe *one-shot learning* și *rețele convoluționale siameze* pentru a face față acestei sarcini. Sistemul analizează reprezentări grafice ale proteinelor pentru a le grupa pe baza asemănărilor structurale ale acestora. Rezultatele experimentale evidențiază faptul că $CVSimP$ depășește, în termeni de F - *measure*, lucrări similare din literatura de specialitate.

Secțiunea 3.1 este structurată după cum urmează. Prezentăm metoda din secțiunea 3.1.1 și rezultatele obținute în secțiunea 3.1.2.

Secțiunea 3.2 este structurată după cum urmează. Discutăm o soluție pentru un set de date open source pentru recunoașterea fructelor este prezentată în secțiunea 3.2.1. Metoda se bazează pe modele neuronale ușoare și, așa cum se arată în secțiunea 3.2.2, realizează performanțe de ultimă generație atât pentru timpul de execuție, cât și pentru performanța de clasificare.

Secțiunea 3.3 prezintă o tehnică de localizare bazată pe *rețea neuronală convoluțională* și este structurată după cum urmează. Prezentăm abordarea în secțiunea 3.3.1 și apoi ilustrăm și comparăm rezultatele obținute în secțiunea 3.3.2.

Secțiunea 3.4 este structurată după cum urmează. În secțiunea 3.4.1 descriem abordarea noastră. Vom acoperi aspecte privind topologia modelului, funcția de cost utilizată și evaluarea. După aceea, în secțiunea 3.4.2 vom descrie setul de date utilizat și setările de parametri ale experimentului. Arătăm și câteva rezultate și facem o discuție legată de un studiu anterior. În cele din urmă, concluziile și ideile viitoare de îmbunătățire sunt discutate în secțiunea 3.4.3.

3.1 Detecția semnăturilor false folosind rețele neuronale convoluționale

În această secțiune descriem un experiment pentru analiza autenticității semnăturilor scrise de mână. Metodele și experimentele descrise în această secțiune au fost introduse în lucrarea originală [Tel17].

Scopul acestei secțiuni este de a prezenta o abordare *machine learning* propusă pentru această clasificare binară, pentru a evidenția performanța sa prin testarea acesteia pe date noi, nevăzute.

3.1.1 Abordarea propusă

Considerăm că această problemă este rezolvată într-o manieră supervizată, deoarece putem folosi seturi de date deja adnotate cu imagini de semnături. În acest scenariu de învățare, modelul va învăța să detecteze dacă o imagine conține o semnătură autentică sau una falsă, analizând astfel de exemple deja adnotate.

Abordarea noastră constă în trei pași. În primul rând, un pas *extragerea caracteristicii* este aplicat asupra datelor de intrare. Pentru acest pas, folosim modelul *inception* pre-antrenat [SVI+15]. Acest model a fost dezvoltat de *Google* și reprezintă o rețea neuronală convoluțională foarte complexă, care este compusă din 59 de straturi. Modelul a fost instruit pe un set considerabil de imagini și a fost capabil să obțină acuratețe de ultimă generație pe probleme foarte complexe, cum ar fi clasificarea *ImageNet* [SVI+15].

Următorul pas constă în instruirea unui clasificator asupra datelor pre-procesate. Mai precis, folosind caracteristicile extrase din setul nostru de date de semnături, ne propunem să construim un clasificator care să învețe să identifice semnăturile falsificate pe baza acestor date de intrare. O mașină cu suport vetcorial (SVM) va fi utilizată pentru a discrimina între semnăturile originale și cele false. SVM va fi apoi *testat* pentru a evalua performanța acestuia.

3.1.1.1 Antrenare

Pe setul de trăsături extrase, este antrenat un *SVM*. Pentru a face acest lucru, luăm toate eșantioanele disponibile din setul de date și aplicăm extractorul de trăsături. Mai mult, setul de instanțe obținut (vectori de caracteristici) este împărțit în 2 seturi: instruire și testare.

Pentru a antrena modelul, sunt utilizați mai mulți hiperparametri, cum ar fi C, funcția kernel, parametrii kernelului (de exemplu, γ pentru RBF). Pentru optimizarea hiperparametrelor, se efectuează o căutare pentru a găsi cele mai potrivite valori.

3.1.1.2 Testare

Performanța modelului *SVM* va fi testată pe un set de testare complet deconectat din setul de date de instruire. Faza de testare va fi efectuată pe date nevăzute.

Deoarece problema considerată este una de clasificare binară, se va calcula *matricea de confuzie*. Pentru crearea matricei de confuzie și calcularea măsurilor, considerăm că semnăturile *false* reprezintă clasa *pozitiv* în timp ce clasa *negativă* este reprezentată de cele *originale*. Un număr mare de valori diferite ale performanței pot fi calculate din matricea confuziei.

Raportăm atât măsurile legate de *acuratețe*, cât și de *matricea de confuzie*, deoarece procedând astfel putem interpreta mai ușor performanța modelului. Mai mult, deoarece setul de testare este dezechilibrat, aceste măsuri pot fi considerate foarte importante.

3.1.2 Set de date și rezultate

3.1.2.1 Set de date

Setul de date utilizat în experimentele noastre este disponibil public [KC13]. Este format din probe adnotate de 4000 de semnături, din care 800 sunt falsuri. Pentru a construi setul de date, mai multe persoane au fost solicitate să-și scrie propria semnătură. Mai mult, a fost solicitată unei alte persoane să încerce să reproducă semnătura originală.

În setul de date avem mai mulți semnatari, fiecare având semnătura originală și unele falsuri. Intenționăm să ne instruim modelul pentru a distinge cele două tipuri de semnătura, fals și original.

3.1.2.2 Rezultate

Pentru experimentele noastre, am folosit implementarea *scikit-learn* a SVM [PVG+11]. 80 % din setul de date a fost rezervat instruirii. Metodologia de testare a fost aplicată folosind restul setului de date.

Pentru experimentul nostru, intervalul de încredere de 95% raportat pentru *acuratețe* pe setul de testare este **[0.935, 0.966]**.

Măsura AUC calculată pentru clasificatorul nostru este **0.92** iar *F-measure* este **0.88**. Aceste valori exprimă o performanță foarte bună pentru modelul de clasificare propus.

Setul de date a fost reorganizat pentru a repeta divizarea aleatorie pentru seturile de antrenament și testare. Experimentul propus a fost repetat de 20 de ori pentru a analiza evoluția măsurii AUC.

Dacă ne uităm numai la măsura de performanță a abordărilor descrise în tabelul 3.1, observăm că abordarea noastră este comparabilă cu lucrările similare. Mai mult, cele interval de încredere obținut prin abordarea noastră este restrâns, în comparație cu cel de la [RGSK11], iar acest lucru dovedește din nou performanța modelului nostru.

#	Metodă	Performanță
1	**Metoda noastră**	$95\% \pm 0.015$
2	Analiză statistică [KC10b, KC10a]	89%
3	Machine learning[SSB06]	84%
4	Deep learning[RGSK11]	$85.03\% \pm 14.25$

Table 3.1: Comparţie cu abordări similare din literatură.

3.1.3 Concluzii

În această lucrare am prezentat o metodă *machine learning* bazată pe un extractor de caracteristici care poate fi utilizat cu succes în rezolvarea problemei de verificare a semnăturii. Având în vedere rezultatele bune, am putea spune că am confirmat din nou că această sarcină complexă este potrivită pentru rezolvarea folosind *machine learning*.

Lucrările ulterioare constau în extinderea experimentului pe mai multe seturi de date de referință, pentru a avea o imagine mai bună asupra capacității metodei propuse. Construirea unei rețele neuronale convoluționale va fi de asemenea luată în considerare.

3.2 Modele redar trebui schimabt uşoare uşoare de învățare profundă pentru recunoașterea fructelor

Scopul acestei secțiuni este de a oferi o soluție eficientă și precisă pentru o problemă de clasificare a imaginilor. Mai exact, ne vom concentra pe o problemă specifică, și anume recunoașterea fructelor. Metodele și experimentele descrise în această secțiune au fost introduse în lucrarea originală [TD19].

Vom studia problema recunoașterii fructelor descrisă în [MO18] cu scopul general de a dezvolta un model de ultimă generație. Pe de altă parte, dorim să abordăm și problematica vitezei de execuție, așa că ne vom concentra exclusiv pe modele ușoare. În sfârșit, intenționăm să purtăm o discuție cu privire la setul de date și să oferim câteva direcții viitoare pentru problema discutată.

Următoarele întrebări de cercetare vor fi cercetate în această secțiune:

RQ1 Cum să proiectăm un model ușor de învățare profundă pentru îmbunătățirea performanței de ultimă generație în recunoașterea fructelor din imagini?

RQ2 Cum se compară modelul de învățare propus în termeni de performanță și timp de execuție cu modelele de ultimă generație?

3.2.1 Metoda propusă

Soluția noastră la problema recunoașterii fructelor este discutată în această secțiune, cu scopul de a răspunde la întrebarea de cercetare RQ1.

3.2.1.1 Set de date

Pentru experimentele noastre, folosim setul de date *Fruit-360* [1] descris în [MO18]. Setul de date constă din mii de imagini cu fructe. Autorii propun 81 de clase de fructe (de exemplu măr roșu, portocală, guavă, prună, zmeură etc).

[1] accesat în noiembrie 2018

3.2.1.2 Metodologie

Să luăm în considerare $I = \{i_1, i_2, \ldots, i_n\}$ un set de imagini reprezentând fructe și $C = \{c_1, \ldots, c_k\}$ un set de clase de fructe (de exemplu, caisă, banană). Astfel, problema noastră de clasificare este formalizată ca fiind aproximarea, din setul de date de instruire I, o funcție țintă $f : I \to C$ care mapează instanțele de la I la clase de la C. Aproximarea învățată hf se numește *ipoteză*.

Pentru rezolvarea problemei, proiectăm variante de rețele neuronale convoluționale. Aceste rețele neuronale sunt capabile să ia o imagine a unui fruct ca date de intrare, să extragă caracteristici și să prezică tipul acestuia.

Modelele noastre sunt compuse din două părți principale: un extractor (coloana vertebrală a rețelei) și stratul de ieșire. Pentru a construi extractorul de caracteristici, adaptăm modelele de ultimă generație în ceea ce privește viteza: *MobileNet V2* [SHZ+18] și *ShuffleNet V2* [MZZS18].

După extractorul de caracteristici, se aplică un strat ce calculează o medie globală [LCY13]. În cele din urmă, se folosește un strat de ieșire compus dintr-un vector de 81 de numere normalizate de o funcție softmax [GBC16].

O altă abordare comună pe care o considerăm interesantă și de obicei foarte utilă este *transfer transfer*. Într-o metodologie de învățare a transferului [GBC16] se poate folosi un model preantrenat pentru a construi un model care să fie potrivit noului set de date.

3.2.1.3 Augmentarea setului de date

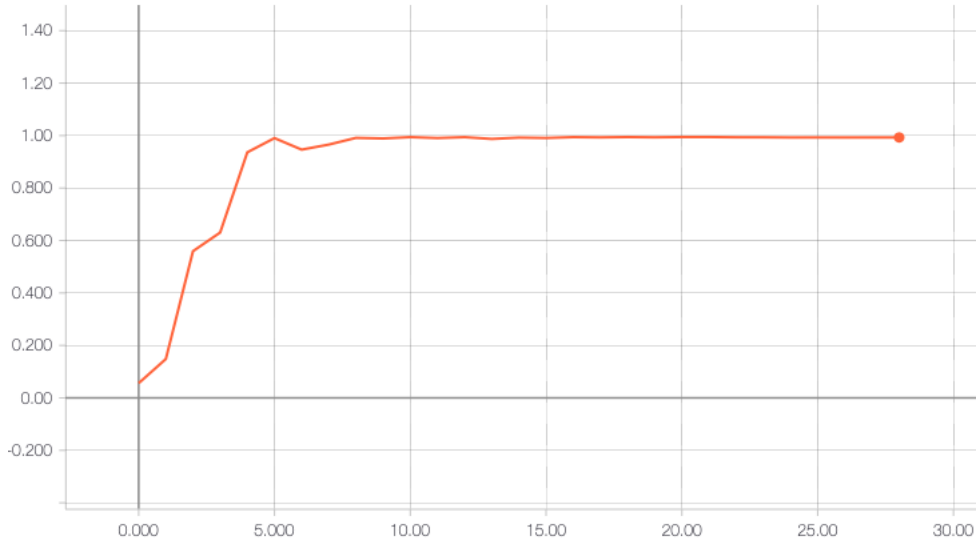


Figure 3.1: Evoluția acurateții pe setul devalidare pentru un experiment cu *MobileNet V2* antrenat de la zero.

Pentru a îmbunătăți performanța modelului, folosim câteva tehnici de augmentare a datelor. Pentru problema noastră, folosim doar flipuri orizontale și verticale, deoarece am observat că toate eșantioanele din setul de date sunt centrate.

3.2.2 Evaluare experimentală

În această secțiune sunt prezentate rezultatele obținute de modelele noastre. Mai mult, secțiunea are scopul de a compara metoda noastră cu lucrări similare și de a arăta avantajele sale în ceea ce privește viteza și precizia.

Tip de model	Transfer learning	Augmentări	Acuratețea pe test	Cea mai bună acuratețe pe test	Cea mai slabă acuratețe pe test
MobileNet V2	Nu	Nu	$97.3\% \pm 0.3$	98.0%	96.4%
MobileNet V2	Nu	Da	$98.0\% \pm 0.2$	98.5%	97.4%
MobileNet V2	ImageNet	Nu	$98.6\% \pm 0.2$	98.9%	97.9%
MobileNet V2	ImageNet	Da	$98.7\% \pm 0.1$	99.1%	98.2%
ShuffleNet V2	Nu	Nu	$97.6\% \pm 0.3$	98.2%	96.9%
ShuffleNet V2	Nu	Da	$98.4\% \pm 0.1$	98.8%	98.1%

Table 3.2: Rezultatele noastre pentru diverse scenarii. Sunt folosite intervale de încredere de 95%.

3.2.2.1 Rezultate și analiză

Efectuăm mai multe tipuri de experimente pentru a găsi cel mai performant model. Diferite setări pentru experimentului sunt combinate pentru a găsi cea mai potrivită metodologie. De exemplu, dorim să evidențiem impactul augmentărilor geometrice (adică flipuri orizontale și verticale aleatorii).

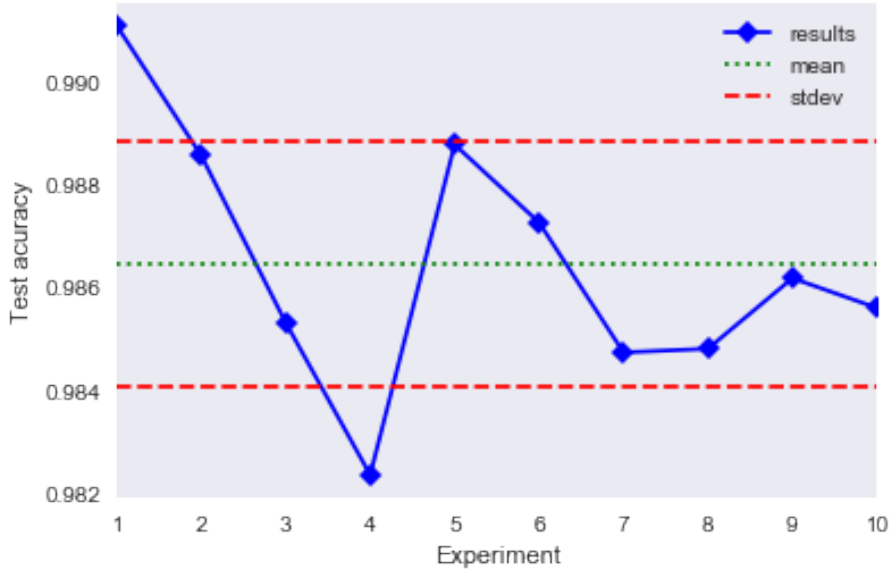


Figure 3.2: Acurateți pe test considerând 10 experimente cu *MobileNet V2*.

În figura 3.1 prezentăm evoluția acurateții în faza de validare pentru un anumit experiment rulat pe *MobileNet V2* care este antrenat de la zero. Nu se utilizează nicio augmentare pentru acest experiment. Observăm că modelul reușește să obțină o precizie de validare bună relativ rapid.

3.2.2.2 Comparație cu lucrări similare

Pentru a răspunde la întrebarea de cercetare RQ2, comparăm modelul de învățare propus în secțiunea 3.2.1 cu modelele de ultimă generație în recunoașterea fructelor din imagini. Comparația se realizează în termeni de precizie a modelului și timp de execuție.

Rezultatele prezentate în tabelul 3.3 arată că modelul nostru depășește orice altă lucrare existentă. De menționat că ceilalți autori au folosit o versiune anterioară a setului de date care avea mai puține clase. De exemplu, rezultatele incluse în [AJW18] sunt raportate pe o

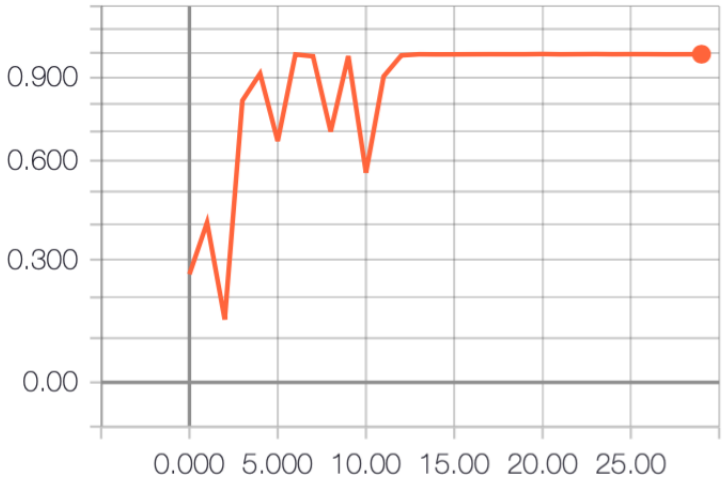


Figure 3.3: Evoluția exactității validării pentru un anumit experiment rulat pe *MobileNet V2* care este instruit folosind inițializarea ImageNet și creșterea datelor.

versiune cu clase de 74 . După cum se menționează în [MO18], setul de date este actualizat continuu.

Model	**Acuratețe pe testy**
Mureşan and Oltean [MO18]	96.3%
Baryła [Bar18]	90.1%
Andersson et al. [AJW18]	96.7%
Ours	$98.7\% \pm 0.1$

Table 3.3: Compararea acurateții obținute cu lucrări existente

3.2.3 Concluzii

În acest studiu am studiat problema clasificării fructelor folosind un set de date de referință. Am realizat studiul cu succes pentru a propune modele ușoare pentru această problemă. Mai mult, așa cum se arată în secțiunea 3.2.2, modelele folosite au obținut performanțe de ultimă generație, atât în ceea ce privește acuratețea clasificării, cât și timpul de execuție. Cel mai important succes al studiului nostru este că ne putem folosi cu ușurință modelele noastre pentru construirea unei aplicații mobile rapide și precise.

Mai mult, în timpul revizuirii din secțiunea 3.2.1.1 am făcut câteva observații care pot fi utile pentru extinderea și îmbunătățirea setului de date. Chiar dacă performanța de test obținută de modelul nostru este suficient de bună pentru stadiul actual, merită menționat faptul că metodologia noastră poate fi extinsă. S-ar putea încerca alți parametrii de experiment, cum ar fi diferite valori de multiplicare a rețelei. Pe de altă parte, ar fi util să avem în vedere o versiune preantrenată *ShuffleNet V2*.

3.3 O tehnică de detectare a documentelor care utilizează rețele neuronale convoluționale pentru sisteme optice de recunoaștere a caracterelor

În secțiunea actuală, introducem o metodă nouă de preprocesare a documentelor pentru recunoașterea optică a caracterelor. Metodele și experimentele descrise în această secțiune

au fost introduse în lucrarea originală [DT19].

Discutăm o metodă nouă de preprocesare bazată pe detectarea documentelor care folosește *învățare profundă* și transformare proiectivă. Metoda foloseste o retea neuronală convoluționala pentru a detecta punctele cheie ale documentului, apoi foloseste aceste puncte pentru proiectarea documentului intr-o forma dreptunghiulara. Mai mult, arătăm că metoda noastră este capabilă atât să detecteze cât și să îndrepte imaginile documentelor.

Capacitatea *rețele neuronale convoluționale* de detecta obiecte, încadrând sarcina de detectare a obiectelor ca o problemă de regresie a fost deja demonstrată cu Overfeat [SEZ+13]. Tehnica noastră folosește această capacitate pentru a prezice valorile (x, y, $width$, $height$) și să o adapteze pentru a prezice mai multe puncte 2D în spațiul imaginii.

3.3.1 Metoda propusă

Prezentăm în această secțiune metodologia noastră pentru modelarea și rezolvarea problemei detectării și corectării grevelor de document într-o manieră de învățare supervizată.

3.3.1.1 Modelul propus

Întrucât încercăm să rezolvăm o problemă de procesare a imaginii folosind învățarea automată, alegem să folosim o *rețea neuronală convoluțională*. Vom construi modelul nostru pe baza *MobileNet* [HZC+17].

Datele de intrare pentru model sunt reprezentate de o fotografie a chitanței. Imaginea este codată ca tensor RGB notat ca x. Având în vedere x, modelul calculează locația a 4 puncte în domeniul imaginii reprezentând colțurile chitanței. Fiecare punct este reprezentat de o pereche notată ca $(\widehat{y}_1, \widehat{y}_2)$ făcând modelul nostru să returneze valori de 8 : $< \widehat{y}_{1,1},\ \widehat{y}_{1,2},\ \widehat{y}_{2,1},\ \widehat{y}_{2,2},\ \widehat{y}_{3,1},\ \widehat{y}_{3,2},\ \widehat{y}_{4,1},\ \widehat{y}_{4,2} >$.

3.3.2 Evaluare experimentală

În această secțiune prezentăm experimentele efectuate pentru a evalua eficacitatea abordării propuse. Este compus din două părți principale: *antrenare* și *testare*.

3.3.2.1 Set de date

Setul de date utilizat în experimentul nostru constă într-o colecție de fotografii ale diferitelor tipuri de încasări în numerar colectate din surse diferite. Toate imaginile au aceeași rezoluție de înaltă calitate (1920 x 1080). Pentru a reduce complexitatea modelului din motive de eficiență în timp și fără a pierde performanța, toate imaginile (și etichetele corespunzătoare) sunt reduse la scară până la 480 x 270 .

Valorile de intrare sunt redimensionate la $[-1, 1]$ [HZC+17]. Pentru experimentul nostru, folosim un set de date compus din 6000 de imagini.

Pentru construirea setului de adnotări am dezvoltat o aplicație web simplă, care a fost folosită pentru a adnota manual setul de date. Aplicația prezintă utilizatorului mostre de fotografii cărora li se cere să marcheze punctele cheie ale documentului. După marcarea coordonatelor, utilizatorul este capabil să vizualizeze imaginea transformată obținută prin aplicarea *colinierii proiective* într-o formă dreptunghiulară folosind punctele cheie pe care le-au plasat.

3.3.2.2 Experimente și rezultate

The *training* is performed on 90% of the dataset. We employ a distinct validation set formed of the remaining 10%.

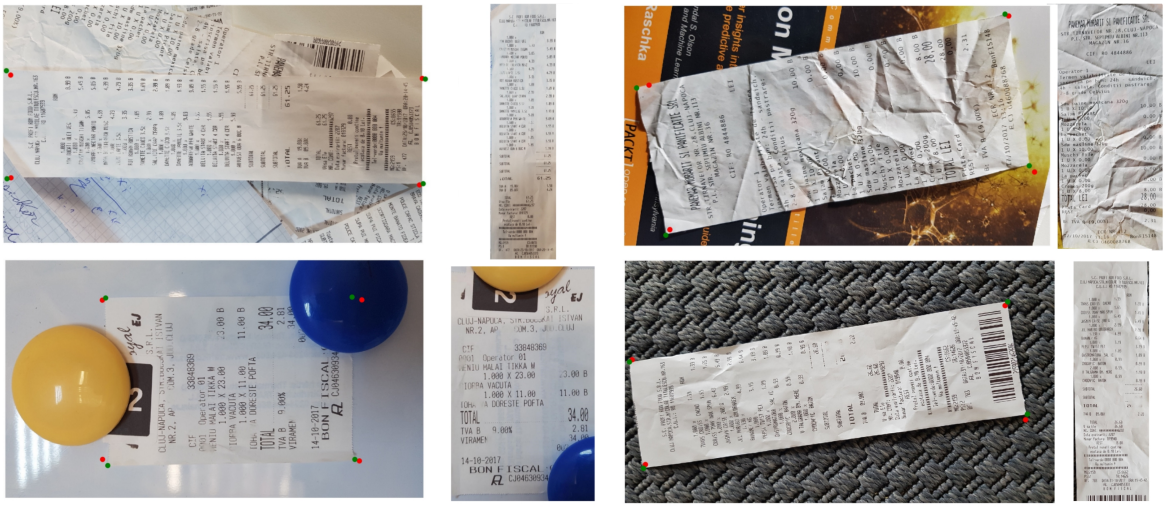


Figure 3.4: Date de testare: imaginea originală și proiecția chitanței detectate. Predicția este marcată de puncte roșii, iar adnotarea este marcată cu puncte verzi. De menționat este faptul că modelul a marcat punctele de referință corect, chiar dacă colțul documentului a fost blocat.

Modelul obținut este testat pe o nouă colecție de imagini, setul de testare. Acesta este format din imagini de 700 de chitanțe în numerar. Încasările în numerar provin de la furnizori diferiți decât cei găsiți în setul de instruire și au fost concepute pentru a fi foarte dificile. Pentru fiecare instanță detectăm cele 4 puncte cheie de și raportăm *eroare medie absolută*, *eroare unghiulară* și *valoare absolută* a unghiului de inclinare pe proiecția rezultată. Efectuăm trei experimente: o versiune utilizează clasicele MSE ca funcție eroare, în timp ce celelalte două versiuni folosesc funcția de eroare propusă. Rezultatele obținute sunt prezentate în tabelul 3.4. Raportăm MAE, eroarea unghiulară și media valorilor absolute ale unghiurilor variabile. Experimentele au fost repetate de 10 de ori pentru fiecare $\lambda \in \{0, 1, 5\}$ pentru a calcula intervalele de încredere de 95% (CI). Datele de testare, inclusiv detectarea și proiecția, sunt reprezentate în Figure 3.4. Valorile raportate arată că cel mai bun model a fost obținut folosind modelul care a utilizat funcția de eroare propusă cu $\lambda = 5$.

Model	MAE	Eroarea unghiulară	Hough
MobileNet $\lambda = 0$	3.66 ± 0.07	4.87 ± 0.07	1.04 ± 0.01
MobileNet $\lambda = 1$	3.37 ± 0.04	4.05 ± 0.12	0.96 ± 0.01
MobileNet $\lambda = 5$	3.38 ± 0.05	3.22 ± 0.14	0.88 ± 0.01

Table 3.4: Rezultate cu intervale de încredere aferente.

Tabelul 3.5 prezintă rezultatele obținute prin compararea metodei noastre cu alte lucrări similare. Pentru o comparație exactă, rulăm toți algoritmii pe același set de testare, calculând unghiul de îndreptare folosind Hough pentru proiecțiile obținute. Rezultatele arată o performanță generală mai bună pentru metoda noastră în comparație cu abordările de la [Xio16] și [JS17].

Durata medie de inferență pe cadru pentru modelul nostru este de 109 ms pe un dispozitiv mobil OnePlus5, folosind TFLite (din Tensorflow 1.12) cu 8 fire de execuție.

Model	Min	Max	Medie	Std
MobileNet $\lambda = 0$	0.00 ± 0	9.61 ± 0.08	1.04 ± 0.01	1.16 ± 0.02
MobileNet $\lambda = 1$	0.00 ± 0	9.33 ± 0.07	0.96 ± 0.01	1.10 ± 0.02
MobileNet $\lambda = 5$	0.00 ± 0	9.62 ± 0.05	0.88 ± 0.01	2.20 ± 0.02
Xiong [Xio16]	0.00	9.45	1.50	1.84
Javed and Shafait [JS17]	0.00	9.79	3.75	2.80

Table 3.5: Comparație cu lucrări similare bazată pe valorile Hough.

3.3.3 Concluzii

Am prezentat o nouă tehnică de preprocesare eficientă pentru a face imaginile mai accesibile pentru algoritmii *OCR*. Principalul beneficiu al tehnicii noastre este adaptabilitatea la situații nevăzute. Procesul combină doi pași foarte importanți pentru obținerea unor rezultate bune: detectarea documentelor și îndreptare.

Lucrări suplimentare vor fi efectuate pentru extinderea metodei noastre pentru a fi utilizate pentru diferite tipuri de documente. Planificăm să generalizăm eroarea unghiulară pentru n puncte cheie, potrivindu-se problemelor de detecție în cazul în care obiectele de interes care apar în forme poligonale de ordin superior.

3.4 *CVSimP*: An approach for predicting proteins' structural similarity using one-shot learning

În acest studiu este abordată problema clasificării proteinelor în funcție de similaritatea lor, din perspectiva computer vision. Vom introduce o nouă abordare *CVSimP* pentru a prezice similaritatea structurală a proteinelor folosind *one-shot learning*. Metodele și experimentele descrise în această secțiune au fost introduse în lucrarea originală [TC20].

În literatura de specialitate *deep learning*, *one-shot learning* este o metodologie pentru a rezolva problemele de clasificare a obiectelor. Majoritatea aplicațiilor sunt legate de computer vision. În timp ce alte metode de clasificare bazată pe învățare automată necesită o antrenare pe mii de eșantioane, modelele bazate pe *one-shot learning* pot fi instruite folosind mai puține date de antrenament. Aceste modele sunt dezvoltate folosind *rețele neuronale siamese* care pot prezice dacă două instanțe aparțin sau nu aceleiași categorii.

Rețelele neuronale siameze au fost inițial propuse de Bromley et al. [BGL+94]. Autorii au sugerat această arhitectură nouă a rețelelor neuronale pentru discriminarea imaginilor cu semnături pe baza autorului lor. O rețea neurală siameză este compusă din două rețele care acceptă două intrări diferite, rețelele respective fiind unite în partea de sus.

Pe scurt, în această lucrare încercăm să răspundem la următoarele întrebări de cercetare:

RQ1 Cum se introduce o abordare *one-shot learning* pentru a prezice similaritatea structurală a proteinelor din imaginile proteice? În acest sens, va fi introdusă o nouă abordare *CVSimP* și validată empiric.

RQ2 Cum se compară abordarea *one-shot learning* propusă în această lucrare cu lucrările similare existente pentru detectarea asemănării structurale inter-proteine?

3.4.1 Metodologie

Cu scopul de a răspunde la prima noastră întrebare de cercetare RQ1, introducem în această secțiune metodologia pe care o folosim pentru rezolvarea problemei de clasificare. Vom defini problema și vom descrie modelul nostru de învățare automată. Acoperim, de asemenea, aspecte privind instruirea și evaluarea performanței.

Vom folosi reprezentările proteice brute. Considerăm că aceste reprezentări brute pot avea o performanță mai bună pentru clasificatorul nostru convoluțional.

Problema poate fi privită ca o sarcină de clasificare binară. Vom considera perechi de imagini proteice drept input cu scopul de a prezice dacă acestea reprezintă aceeași proteină sau nu.

Model

Modelul nostru este o rețea neuronală convoluțională siameză. Arhitectura modelului este prezentată mai jos:

$$
\begin{aligned}
&INPUT(100x100x3) \rightarrow \\
&\rightarrow CONV(3x3, 6) \rightarrow ReLU \rightarrow BatchNORM \rightarrow \\
&\rightarrow [CONV(3x3, 8) \rightarrow ReLU \rightarrow BatchNORM] \cdot 2 \rightarrow \\
&\rightarrow FC(500) \rightarrow ReLU \rightarrow FC(250) \rightarrow ReLU \rightarrow OUT(8)
\end{aligned}
$$

Funcția de cost

Pentru instruirea modelului, având ca obiectiv principal optimizarea vectorilor de ieșire codificați, folosim *contrastive loss* [HCL06]. Funcția a fost propusă de Hadsell et al. și reprezintă o metodă de a învăța modelul să învețe mapări astfel încât eșantioanele marcate ca fiind similare să fie trase împreună, în timp ce cele marcate care nu sunt similare sunt îndepărtate. Funcția folosește distanța euclidiană care poate fi definită ca norma L_2 a diferenței dintre cei doi vectori.

3.4.2 Rezultate experimentale și discuție

În această secțiune descriem studiul nostru de caz asupra unui set de date de imagini obținute din *protein data bank* [BWF+00, BHN04], urmărind să răspundem în continuare la întrebările de cercetare RQ1 și RQ2. Metodologia și rezultatele obținute sunt descrise în continuare. În primul rând, setul de date utilizat în experimentele noastre este descris în Secțiunea 3.4.2. Apoi, vom continua cu prezentarea în secțiunea 3.4.2 experimentele efectuate și în secțiunea 3.4.2 rezultatele obținute.

Set de date

Setul nostru de date constă din imagini cu 57 de proteine. Pentru fiecare proteină, am capturat diverse imagini din unghiuri diferite. Toate acestea au fond negru iar proteina este centrată. Aceste proteine sunt clasificate ca aparținând celor 9 superfamilii, așa cum s-a arătat în tabelul 3.6. Pentru fiecare superfamilie, a treia coloană prezintă proteinele care sunt utilizate pentru antrenarea modelului, iar ultima coloană reprezintă proteina folosită pentru testare.

Am colectat imagini pentru fiecare proteină introdusă în studiul nostru de la *protein data bank* [BWF+00, BHN04]. Reprezentările sunt create folosind *NGL* [RBV+18].

Experimente

Modelul este instruit pentru mai multe epoci pe perechi create folosind reprezentarea grafică a proteinelor din setul de antrenament și este evaluat la sfârșitul fiecărei epoci folosind un set de validare fix. Perechile utilizate pentru antrenament sunt recreate aleatoriu după sfârșitul fiecărei epoci. Pentru dezvoltarea modelului și a metodei de formare și evaluare, am folosit *PyTorch* [PGCC17].

#	**Superfamilie**	**Proteine** folosite pentru antrenament	Proteină de test
1	**3.20.20.80**	{1B1Y, 1CNV, 1ITX, 1JFX, 1KFW, 1NAR, 1VFF, 2EBN}	$1EDG$
2	**1.10.490.10**	{1ITH, 1MBA, 2HBG, 2LHB, 1ASH, 1DLW, 1ECA}	$1HLB$
3	**1.10.238.10**	{1OMR, 1SRA, 2SAS, 1CB1, 1IQ3}	$1UHN$
4	**2.40.50.140**	{1SLJ, 1YVC, 1EOV, 1JT8, 1KRS}	$1AH9$
5	**2.60.120.260**	{1NKG, 1PMJ, 1GUI, 1I5P, 1K45}	$1ULO$
6	**3.30.30.10**	{1PE4, 1SEG, 1BCG, 1GPT, 1I2U}	$1JXC$
7	**2.60.40.10**	{1R6V, 2FCB, 1JBJ, 1JE6, 1NCT}	$1OLL$
8	**3.40.50.150**	{1Y8C, 1DUS, 1F3L, 1YUB}	$1AF7$
9	**2.160.20.10**	{1RU4, 1VBL, 1BHE, 1EE6}	$1QCX$

Table 3.6: Proteine folosite în acest studiu [ATC18].

Metodă	**Măsură de evaluare**	$1EDG$ (F_1)	$1HLB$ (F_2)	$1UHN$ (F_3)	$1AH9$ (F_4)	$1ULO$ (F_5)	$1JXC$ (F_6)	$1OLL$ (F_7)	$1AF7$ (F_8)	$1QCX$ (F_9)	**Per total**
CVSimP	*F-measure*	0.967	0.835	0.828	0.905	0.848	0.901	0.856	0.828	0.858	0.867
Autoencodere [TCC19]	*F-measure*	1	1	0.888	0.923	0.545	1	0.750	0.533	1	0.845
Clustering [TCB19]	*F-measure*	-	-	-	-	-	-	-	-	-	0.715

Table 3.7: Rezultate experimentale $CVSimP$. O comparație cu lucrări similare este prezentată

Am utilizat o abordare de *mini batch SGD*, îmbunătățită cu optimizatorul Adam [KB14] folosind o dimensiune a *batchurilor* de 32. Stabilim rata de învățare inițială la $1e-3$ și folosim o politică de reducere atunci când nu este detectată nicio îmbunătățire pentru 5 epoci. În acest caz, rata de învățare este redusă cu 50%. Se păstrează cel mai performant model din setul de validare.

Rezultate și discuție

În tabelul 3.7 am raportat performanța pentru fiecare proteină. Este marcată de asemenea superfamilia corespunzătoare (așa cum se notează în tabelul 3.6). Scorul total $F1$ obținut este de 0,867.

Pentru comparație cu lucrările anterioare, am avut în vedere două abordări pe care le-am propus și testat recent pe același set de proteine. Abordările s-au bazat pe autoencodere [TCC19] și clustering [TCB19].

3.4.3 Concluzii

Am prezentat în această lucrare o abordare $CVSimP$ bazată pe *rețele neuronale convoluționale siamese* pentru a prezice similaritatea structurală a proteinelor folosind imagini proteice obținute de la *protein data bank* [BWF+00]. A fost utilizat un set de date format din imagini cu 57 de proteine care sunt incluse în nouă superfamilii. Am dezvoltat o abordare de învățare unică pentru care s-au obținut rezultate încurajatoare.

Concluzii

În această teză am investigat problema analizei tranzițiilor conformaționale ale proteinelor, cu scopul mai general de a contribui la o înțelegere cuprinzătoare a problemei. Am prezentat abordările actuale de ultimă generație și am propus câteva perspective computaționale noi asupra problemei, bazate pe *machine learning*.

O altă direcție de cercetare pe care am urmărit-o în studiul nostru a fost *computer vision*. Scopul principal a fost să dezvoltăm metode bazate pe învățare profundă pentru abordarea mai multor probleme din lumea reală.

Am susținut că, deși aceste două domenii sunt distincte, complexitatea care le caracterizează poate fi un punct de legătură. Prin urmare, efortul nostru principal a fost orientat către eficiența tehnicilor de învățare profundă pentru dezvoltarea metodelor de ultimă generație.

Din perspectivă bioinformatică, am subliniat, de asemenea, prin mai multe experimente, că informațiile obținute prin analizarea tranzițiilor conformaționale ale proteinelor captează relațiile dintre proteinele înrudite, relații care sunt confirmate din perspectivă biologică.

Mai mult, am efectuat un studiu privind aplicarea metodelor de învățare nesupervizată pentru analiza tranzițiilor conformaționale proteice pentru a extrage informații despre similaritatea lor structurală. Un studiu privind aplicarea *deep autoencoders* a fost de asemenea dezvoltat pentru a da o mai bună înțelegere a dinamicii proteinelor. Experimentele efectuate au arătat că *autoencoderele* sunt modele eficiente nesupervizate capabile să învețe structura proteinelor, precum și să descopere asemănarea structurală dintre proteine. Mai mult, am obținut o dovadă empirică conform căreia rețelele cu autosupervizare sunt capabile să codifice tipare ascunse relevante dintr-o perspectivă biologică. Abordarea a fost extinsă pentru analiza inter-proteine folosind un ansamblu de autoencodere profunde.

De asemenea, *clustering* a fost folosit ca metodă de *clasificare nesupervizată* pentru investigarea relevanței valorilor RSA pentru a prezice tranzițiile interne ale proteinelor. Metoda a fost folosită și pentru a găsi similitudini între proteine. *Analiza componentelor principale* a fost explorată pentru vizualizarea datelor și utilizată pentru a examina cum evoluează valorile RSA între tranzițiile conformaționale. Experimentele efectuate pe mai multe proteine au evidențiat faptul că valorile RSA se schimbă fără probleme între tranzițiile conformaționale.

Din perspectiva *computer vision* am investigat diverse probleme. Am abordat problema de clasificare a imaginii pe două sarcini particulare care au fost abordate folosind *rețele neuronale convolutive profunde*. Astfel, am prezentat lucrările noastre pentru analiza autenticității semnăturilor. Mai mult, am dezvoltat o tehnică de ultimă generație atât din punct de vedere al performanței de clasificare, cât și din perspectiva eficienței pe un set de date open source pentru clasificarea fructelor. Am discutat și despre o altă sarcină, și anume localizarea documentelor. O metodă pentru această problemă a fost propusă cu obiectivul mai larg de a crea o tehnică de îndreptare a documentelor. Un studiu interdisciplinar în care am folosit tehnici computer vision pentru a studia proteine a fost în final discutat.

Planificăm să extindem analiza rezultatelor analizei proteinelor obținute în această teză din punct de vedere biologic. Pe baza acestui studiu ne propunem să avansăm cercetările noastre pentru a prezice tranzițiile conformației proteice. În plus, viitoarele lucrări vor

fi realizate pentru a combina valorile RSA cu reprezentările structurale ale alfabetului [PFK10] de tranziții proteice pentru a obține o perspectivă suplimentară asupra tranzițiilor conformației proteice.

În ceea ce privește viziunea computerului, intenționăm să continuăm îmbunătățirea performanței noastre actuale cu privire la sarcinile discutate. Mai mult, intenționăm să continuăm investigațiile noastre pentru aplicarea tehnicilor similare pentru domenii practice noi, precum îngrijirea sănătății și probleme legate de mediu, cum ar fi gestionarea și reciclarea deșeurilor.

Bibliografie

[ACT18] Silvana Albert, Gabriela Czibula, and Mihai Teletin. Analyzing the impact of protein representation on mining structural patterns from protein data. In *2018 IEEE 12th International Symposium on Applied Computational Intelligence and Informatics (SACI)*, pages 000533–000538, May 2018.

[AJW18] Joel Andersson, Eskil Jarlskog, and Richard Wang. Fruit recognition. *Report, University of California San Diego*, 2018.

[ATC18] Silvana Albert, Mihai Teletin, and Gabriela Czibula. Analysing protein data using unsupervised learning techniques. *International Journal of Innovative Computing, Information and Control*, 14:861–880, 2018.

[Bar18] Mateusz Baryła. What is this fruit? neural network application for vietnamese fruit recognition. In *ITM Web of Conferences*, volume 20, page 02009. EDP Sciences, 2018.

[BBCG13] Jacques M Bahi, Wojciech Bienia, Nathalie Côté, and Christophe Guyeux. Is protein folding problem really a np-complete one? first investigations. *arXiv preprint arXiv:1306.1372*, 2013.

[BCB94] Daniel Pierre Bovet, Pierluigi Crescenzi, and D Bovet. *Introduction to the Theory of Complexity*. Prentice Hall London, 1994.

[BDF+15] Guillaume Bouvier, Nathan Desdouits, Mathias Ferber, Arnaud Blondel, and Michael Nilges. An automatic tool to analyze and cluster macromolecular conformations based on self-organizing maps. *Bioinformatics*, 31(9):1490–1492, 2015.

[BGL+94] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a" siamese" time delay neural network. In *Advances in neural information processing systems*, pages 737–744, 1994.

[BHN04] Helen Berman, Kim Henrick, and Haruki Nakamura. Berman, h, henrick, k and nakamura, h. announcing the worldwide protein data bank. nat struct biol 10: 980. *Nature structural biology*, 10:980, 01 2004.

[BL98] BONNIE BERGER and TOM LEIGHTON. Protein folding in the hydrophobic-hydrophilic (hp) model is np-complete. *Journal of Computational Biology*, 5(1):27–40, 1998. PMID: 9541869.

[BPC+17] Maria-Iuliana Bocicor, Alesandro Pandini, Gabriela Czibula, Silvana Albert, and Mihai Teletin. Using Computational Intelligence Models for Additional Insight into Protein Structure. *Studia Universitatis "Babeş-Bolyai" Informatica*, 62:107—-119, 2017.

[BWF+00] H.M Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The Protein Data Bank. *Nucleic Acids Res.*, 28:235–242, 2000.

[C+15] F. Chollet et al. Keras. https://github.com/fchollet/keras, 2015.

[CCT19] Gabriela Czibula, Carmina Codre, and Mihai Teletin. *AnomalP*: A new approach for detecting anomalous protein conformations using deep autoencoders. *Expert systems with applications*, page under review, 2019.

[Cho16] François Chollet. Xception: Deep learning with depthwise separable convolutions. *CoRR*, abs/1610.02357, 2016.

[DDS+09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition.*, pages 248–255. IEEE, 2009.

[DT19] Lorand Dobai and Mihai Teletin. A document detection technique using convolutional neural networks for optical character recognition systems. *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning 2019, Bruges, Belgium*, pages 547–552, 2019.

[GBC16] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[HCL06] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006.

[HZC+17] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[ICT17] Vlad-Sebastian Ionescu, Gabriela Czibula, and Mihai Teletin. Supervised learning techniques for body mass estimation in bioarchaeology. In *IEEE 7th International Workshop on Soft Computing Applications (SOFA 2016)*, pages 71–86. Springer, 2017.

[ITV16] Vlad-Sebastian Ionescu, Mihai Teletin, and Estera-Maria Voiculescu. Machine learning techniques for age at death estimation from long bone lengths. In *IEEE 11th International Symposium on Applied Computational Intelligence and Informatics (SACI 2016)*, pages 457 – 462. IEEE Hungary Section, 2016.

[JS17] Khurram Javed and Faisal Shafait. Real-time document localization in natural images by recursive application of a cnn. In *Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on*, volume 1, pages 105–110. IEEE, 2017.

[KB14] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[KC10a] Bence Kovari and Hassan Charaf. Analysis of intra-person variability of features for off-line signature verification. *W. Trans. on Comp.*, 9(11):1359–1368, November 2010.

[KC10b] Bence Kovari and Hassan Charaf. Statistical analysis of signature features with respect to applicability in off-line signature verification. In *Proceedings of the 14th WSEAS International Conference on Computers: Part of the 14th WSEAS CSCC Multiconference - Volume II*, ICCOMP'10, pages 473–478, Stevens Point, Wisconsin, USA, 2010. World Scientific and Engineering Academy and Society (WSEAS).

[KC13] Bence Kovari and Hassan Charaf. A study on the consistency and significance of local features in off-line signature verification. *Pattern Recognition Letters*, https://www.aut.bme.hu/Pages/Research/Signature/Resources, 2013.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[KU03] Daniel Keysers and Walter Unger. Elastic image matching is np-complete. *Pattern Recogn. Lett.*, 24(1):445–453, January 2003.

[LCY13] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *CoRR*, abs/1312.4400, 2013.

[LJN06] D.P. Lewis, T. Jebara, and W. S. Noble. Support vector machine learning from heterogeneous data: an empirical analysis using protein sequence and structure. *Bioinformatics*, 22(22):2753–2760, 2006.

[MDH+10] T. Meyer, M. D'Abramo, A. Hospital, M. Rueda, C. Ferrer-Costa, A. Pérez, O. Carrillo, J. Camps, C. Fenollosa, D. Repchevsky, J.L. Gelpí, and M. Orozco. MoDEL: A database of atomistic molecular dynamics trajectories. *Structure*, 18(11):1399 – 1409, 2010.

[MJC02] K. K. Moon, R. L. Jernigan, and G. S. Chirikjian. Efficient generation of feasible pathways for protein conformational transitions. *Biophysical Journal*, 83(3):1620–1630, 2002.

[MMC08] G. Morra, M. Meli, and G. Colombo. Molecular dynamics simulations of proteins and peptides: from folding to drug design. *Current Protein and Peptide Science*, 9:2181–2196, 2008.

[MO18] Horea Mureşan and Mihai Oltean. Fruit recognition from images using deep learning. *Acta Universitatis Sapientiae, Informatica*, 10(1):26–42, 2018.

[MZZS18] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. *arXiv preprint arXiv:1807.11164*, 2018.

[PFK10] A. Pandini, A. Fornili, and J. Kleinjung. Structural alphabets derived from attractors in conformational space. *BMC Bioinformatics*, 11(97):1–18, 2010.

[PGCC17] Adam Paszke, Sam Gross, Soumith Chintala, and Gregory Chanan. Pytorch: Tensors and dynamic neural networks in python with strong gpu acceleration. https://github.com/pytorch/pytorch, 2017.

[PVG+11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[RBV+18] Alexander S Rose, Anthony R Bradley, Yana Valasatava, Jose M Duarte, Andreas Prlić, and Peter W Rose. Ngl viewer: web-based molecular graphics for large complexes. *Bioinformatics*, 34(21):3755–3758, 2018.

[RGSK11] Bernardete Ribeiro, Ivo Gonçalves, Sérgio Santos, and Alexander Kovacec. Deep learning networks for off-line handwritten signature recognition. In César San Martin and Sang-Woon Kim, editors, *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pages 523–532, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[SEZ+13] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013.

[SHZ+18] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.

[SSB06] Harish Srinivasan, Sargur N. Srihari, and Matthew J. Beal. Machine learning for signature verification. In Prem K. Kalra and Shmuel Peleg, editors, *Computer Vision, Graphics and Image Processing*, pages 761–775, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[SVI+15] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567, 2015.

[TC20] Mihai Teletin and Gabriela Czibula. Cvsimp: An approach for predicting proteins' structural similarity using one-shot learning. In *2020 IEEE 14th International Symposium on Applied Computational Intelligence and Informatics (SACI)*, volume in press, 2020.

[TCAB18] Mihai Teletin, Gabriela Czibula, Silvana Albert, and Mariana-Iuliana Bocicor. Using unsupervised learning methods for enhancing protein structure insight. *Procedia Computer Science*, 126:19 – 28, 2018. Knowledge-Based and Intelligent Information Engineering Systems: Proceedings of the 22nd International Conference, KES-2018, Belgrade, Serbia.

[TCB+18] Mihai Teletin, Gabriela Czibula, Mariana-Iuliana Bocicor, Silvana Albert, and Alessandro Pandini. Deep autoencoders for additional insight into protein dynamics. In *Artificial Neural Networks and Machine Learning – ICANN 2018*, pages 79–89, Cham, 2018. Springer International Publishing.

[TCB19] Mihai Teletin, Gabriela Czibula, and Maria-Iuliana Bocicor. Using clustering models for uncovering proteins' structural similarity. In *2019 IEEE 13th International Symposium on Applied Computational Intelligence and Informatics (SACI)*, volume in press, pages 185–190, 2019.

[TCC19] Mihai Teletin, Gabriela Czibula, and Carmina Codre. Autosimp: An approach for predicting proteins' structural similarities using an ensemble of deep autoencoders. In *International Conference on Knowledge Science, Engineering and Management*, pages 49–54. Springer, 2019.

[TD19] Mihai Teletin and Lorand Dobai. Lightweight models for fruits recognition. In *2019 IEEE 13th International Symposium on Applied Computational Intelligence and Informatics (SACI)*, volume in press, pages 69–74, 2019.

[Tel17] Mihai Teletin. Machine Learning Techniques for Detecting False Signatures. *Studia Universitatis "Babeş-Bolyai" Informatica*, 62:49—-59, 2017.

[TT09] Nobuhiko Tokuriki and Dan S. Tawfik. Protein dynamism and evolvability. *Science*, 324(9524):203–207, 2009.

[Xio16] Ying Xiong. Fast and accurate document detection for scanning, August 2016.

[YG04] Y. Ye and A. Godzik. FATCAT: a web server for flexible structure comparison and structure similarity searching. *Nucleic Acids Res.*, 32:582—-585, 2004.