UNIVERSITATEA DIN
BUCUREȘTI

FACULTATEA DE
MATEMATICĂ ȘI
INFORMATICĂ

SPECIALIZAREA CALCULATOARE
ȘI TEHNOLOGIA INFORMAȚIEI

Lucrare materie

proiect

DETECȚIE INFECȚIE URINARĂ ÎN MICROSCOPIE

Student

Popa Robert-Daniel

București, iunie 2025

Rezumat

Lucrarea de față utilizează un set de date de microscopie clinică pentru dezvoltarea unui model de inteligență artificială capabil să detecteze infecțiile urinare. Modelele realizate efectuează o clasificare binară de tipul „celulă de interes" sau „fundal". Pentru această sarcină au fost utilizate atât metode de învățare nesupravegheată (DBSCAN), cât și metode de învățare supravegheată (U-Net).

În urma analizei exploratoritorie a datelor, s-a constatat că acestea sunt foarte corelate, iar trecerile de la fundal la o celulă de interes pot fi ușor confundate cu trecerile către zgomot de fundal. Cel mai promițător model, un U-Net cu trei straturi de encoder, trei de decoder și un strat bottleneck, a reușit să detecteze în medie 83,1% dintre celulele de interes, cu o rată de false positive de doar 0,28% pentru fundal.

Modelul nesupravegheat a obținut o performanță remarcabilă, detectând în medie 71,2% dintre celulele de interes, cu o rată de false positive de 0,48%, însă timpul de execuție este semnificativ mai mare.

Cuprins

1 Introducere **4**

2 Cuprins **5**

2.1 Explorarea datelor 5

2.2 Învățarea nesupravegheată 5

2.3 Învățarea supravegheată 9

Capitolul 1

Introducere

Lucrarea de față explorează două direcții fundamentale ale învățării automate: învățarea nesupravegheată și învățarea supravegheată. Fiecare dintre aceste abordări oferă perspective diferite în procesul de analiză și procesare a imaginilor celulare, punând în evidență atât avantajele, cât și limitările fiecăreia. Învățarea nesupravegheată a fost utilizată pentru explorarea structurii datelor fără etichete, identificarea formelor și a grupărilor naturale în imagini, în timp ce învățarea supravegheată a permis antrenarea unor modele specializate pentru segmentarea precisă a celulelor, pe baza etichetelor furnizate.

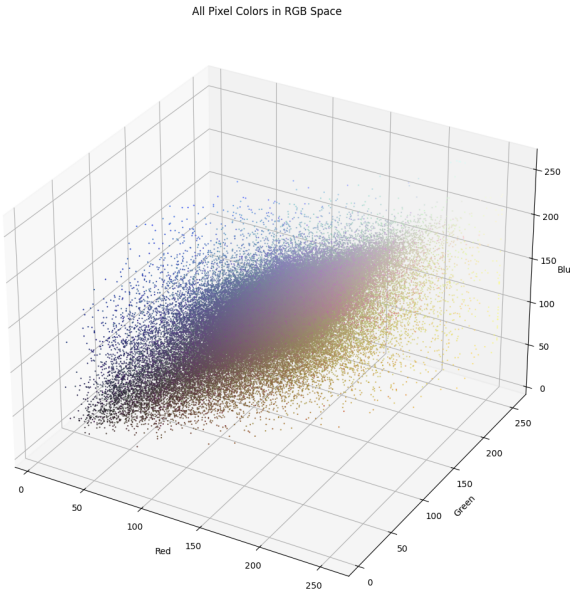
Pe parcursul acestei lucrări, am testat multiple metode și arhitecturi, învățând despre rolul critic al preprocesării, alegerea funcției de pierdere, adaptarea arhitecturii rețelei și selecția metricilor de evaluare.

Capitolul 2

Cuprins

2.1 Explorarea datelor

Explorarea datelor a fost executată în următorul mod am plotat cele 3 canale RGB, dar nimic folositor nu a putut fi extras, așa că am decis pe poate modul în care datele sunt reprezentate este problema așa ca am decis sa folosesc alte moduri de reprezentare, însa și în cadrul acestora același lucru a putut fi sesizat.



2.2 Învățarea nesupravegheată

În faza inițială a proiectului, am testat algoritmul K-Means (utilizat anterior în laborator). Totuși, acesta s-a dovedit inadecvat pentru această sarcină, întrucât necesită specificarea prealabilă a numărului de clustere, ceea ce nu este posibil în cazul imaginilor microscopice, unde numărul de celule variază.

Ca alternativă, am ales algoritmul DBSCAN (Density-Based Spatial Clustering of Applications with Noise), care nu necesită această informație și este capabil să identifice clustere de formă arbitrară. După ajustarea hiperparametrilor (fine-tuning), modelul a reușit să detecteze celulele într-un mod satisfăcător, așa cum se poate observa în Figura 2.3.

Rezultatele au fost convertite într-un format compatibil cu măștile binare din setul de date de antrenare (Figura 2.3). Totuși, modelul a generat un număr considerabil de detecții false (noise), provocate de petele din fundal. Pentru a îmbunătăți precizia, am aplicat un filtru suplimentar pentru a elimina aceste elemente nedorite (Figura 2.7).

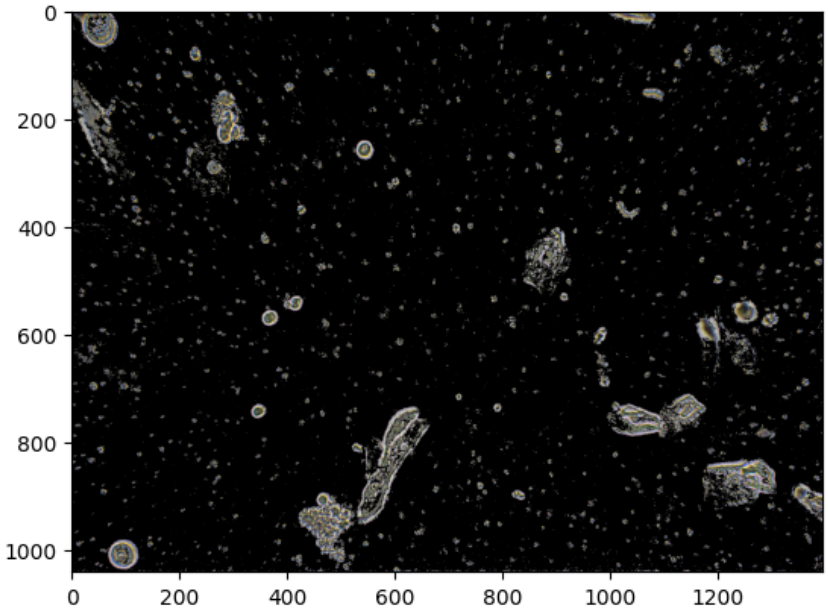


Figura 2.1: Rezultatul inițial al modelului DBSCAN

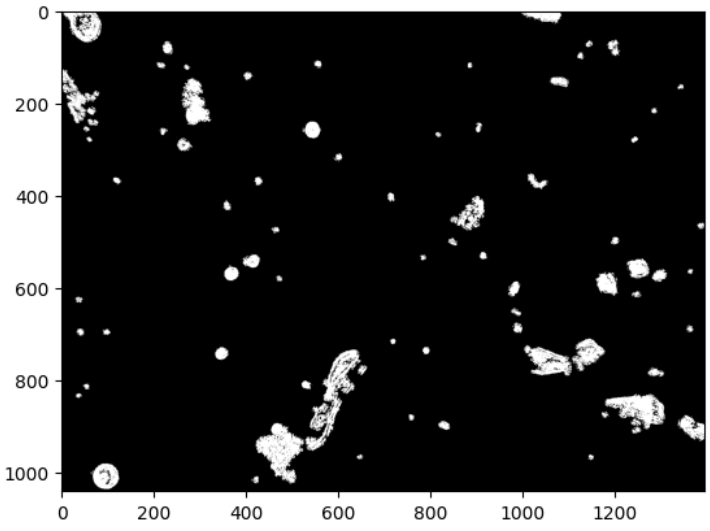


Figura 2.2: Conversia în mască binară

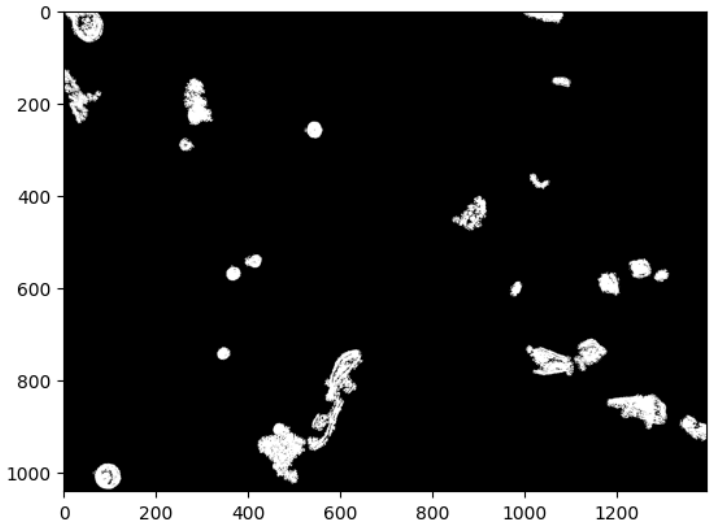


Figura 2.3: Imaginea după filtrarea zgomotului

Pentru a evalua performanța algoritmului DBSCAN, au fost utilizate trei metrici cantitative: **Dice Score**, **Hausdorff Distance** și **Normalized Surface Dice**. Acestea sunt recomandate și în cadrul laboratorului de către domnul laborant Ceaușescu, fiind adecvate pentru evaluarea segmentărilor medicale.

Dice Score măsoară gradul de suprapunere între masca prezisă și cea de referință (ground truth), fiind sensibil la corectitudinea generală a segmentării. **Hausdorff Distance** evaluează diferențele la nivelul marginilor segmentărilor, reflectând cât de bine sunt aliniate contururile detectate. **Normalized Surface Dice** ia în considerare distanța dintre suprafețele segmentărilor, oferind o evaluare mai robustă pentru forme complexe sau neregulate.

În plus față de aceste metrici, este esențială și analiza consistenței detecției în funcție de *tipul de celulă*. Unele tipuri pot apărea în cantitate redusă, pot avea dimensiuni foarte mici sau pot fi dificil de distins față de fundal, ceea ce complică procesul de segmentare. Prin urmare, evaluarea pe categorii ajută la identificarea punctelor forte și a limitărilor metodei utilizate.

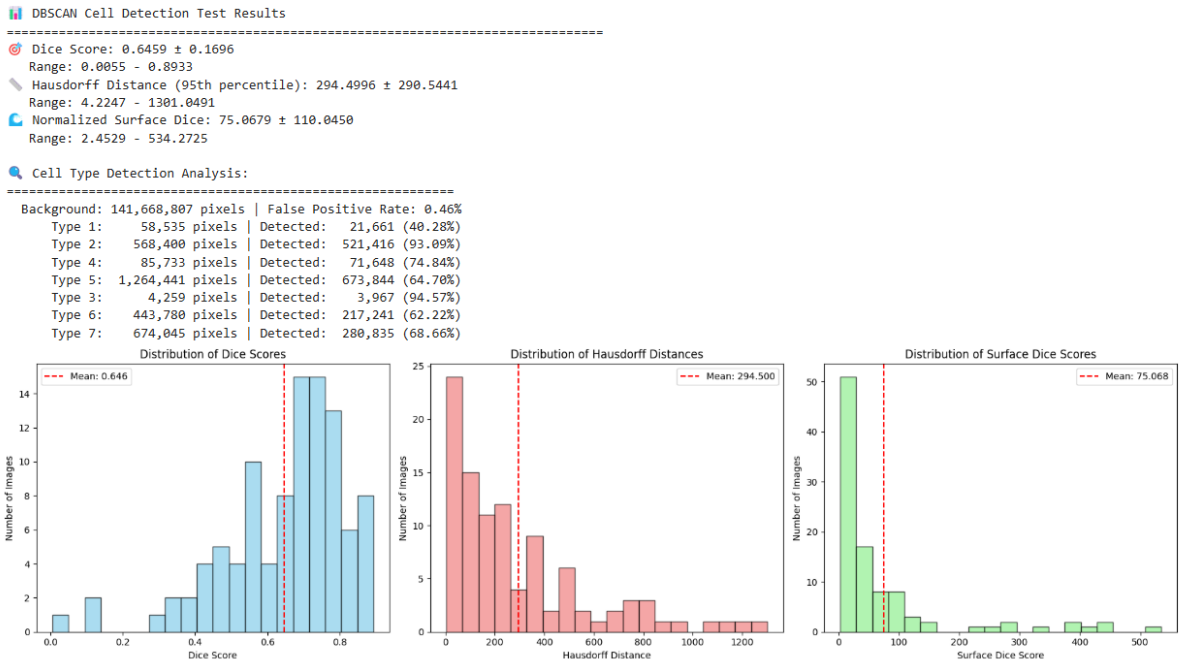


Figura 2.4: Analiza performanței DBSCAN pe baza metricalor cantitative și a tipurilor de celule

Deși în majoritatea cazurilor algoritmul DBSCAN a obținut rezultate rezonabile, au existat și exemple în care segmentarea automată a eșuat semnificativ, evidențând limitările abordării nesupravegheate în prezența unor variații semnificative în aspectul imaginilor sau în structura celulelor.

În Figura 2.3, Figura 2.3 și Figura 2.7 sunt ilustrate trei astfel de exemple, unde valorile Dice Score sunt extrem de scăzute. Aceste valori indică o confuzie între fundal și celule. Astfel de erori pot fi cauzate de celule atipice, zgomot de fundal accentuat sau variații slabe de culoare între regiuni relevante și irelevante.

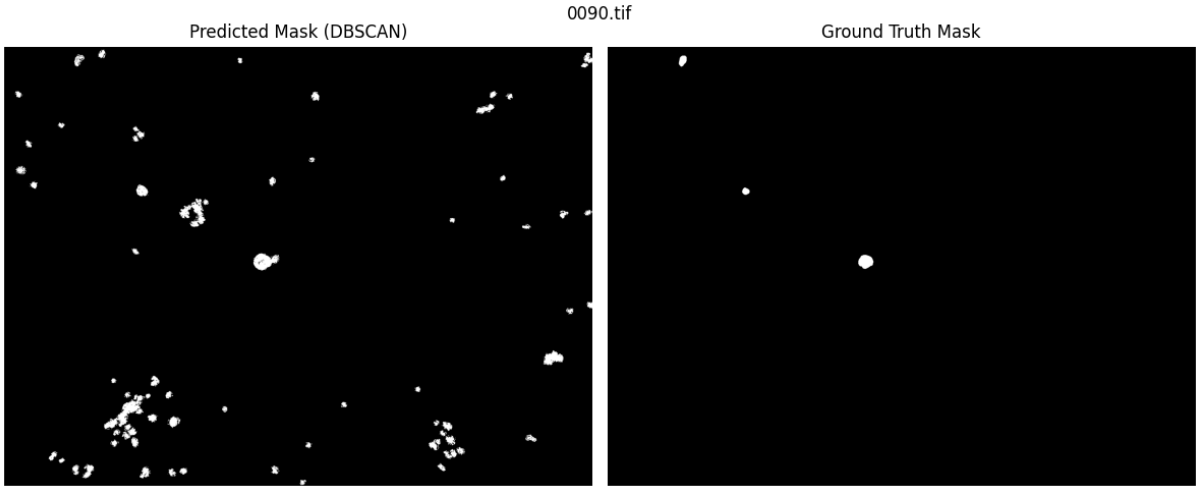


Figura 2.5: Exemplu outliner - Dice Score: 0.1275

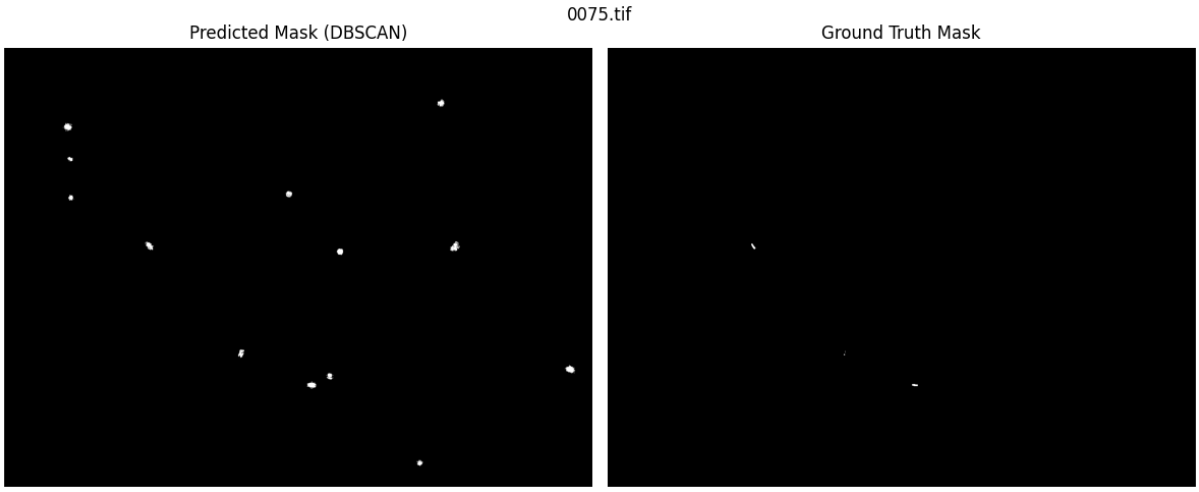


Figura 2.6: Exemplu outliner - Dice Score: 0.1080

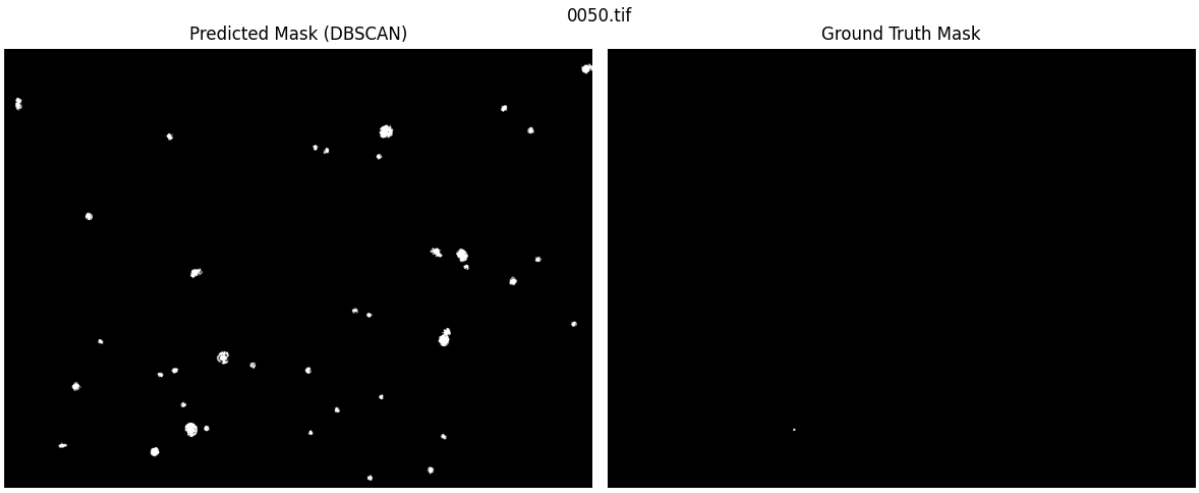


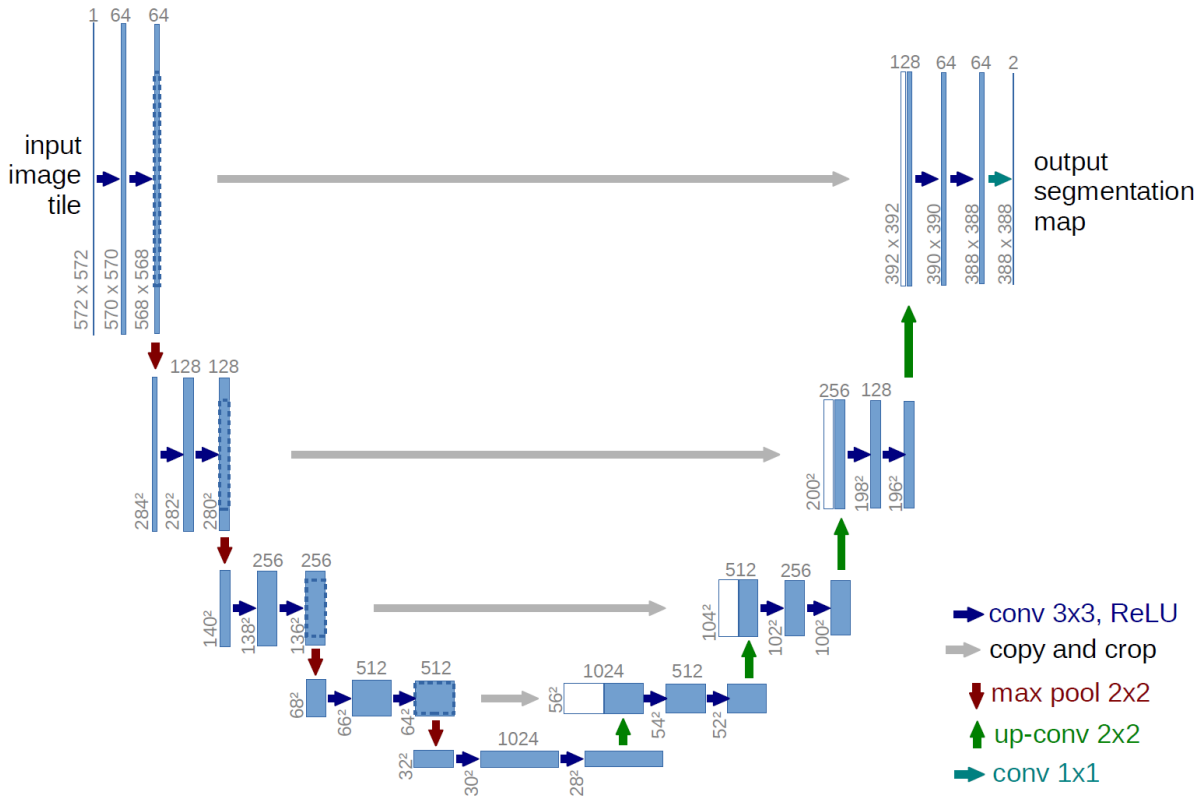
Figura 2.7: Exemplu outliner - Dice Score: 0.0055

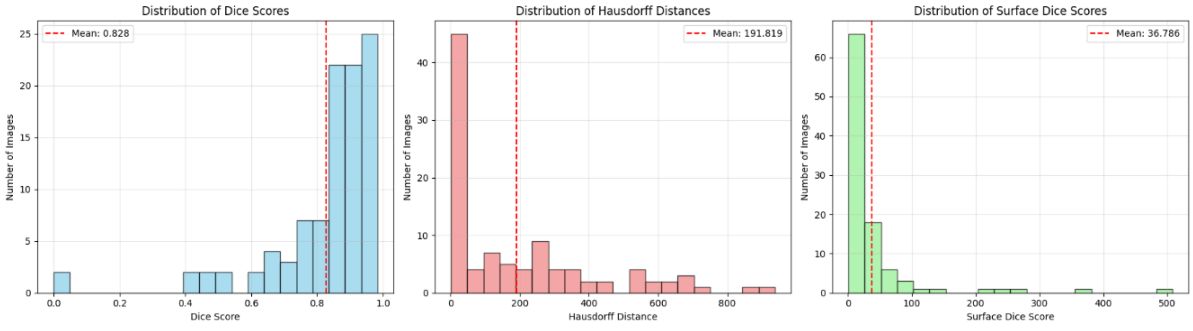
2.3 Învățarea supravegheată

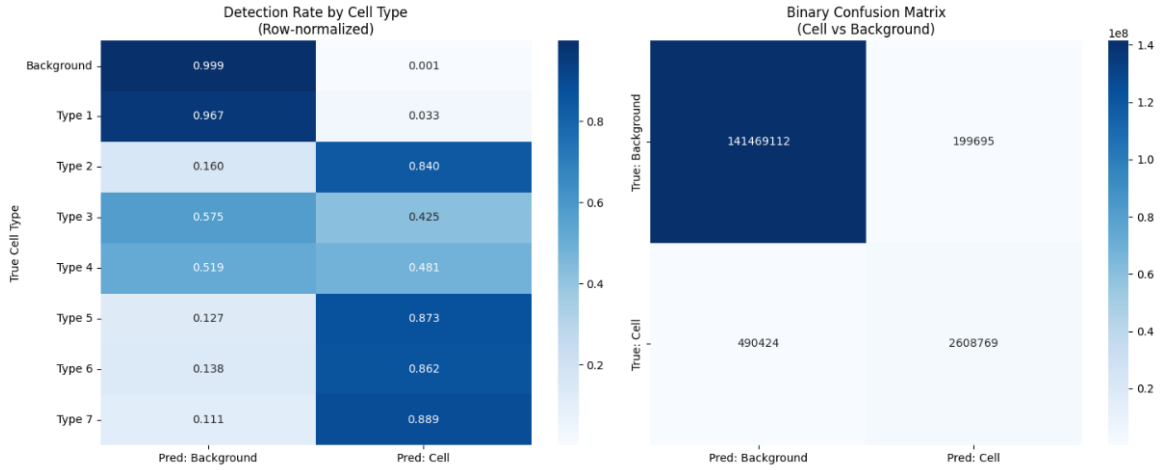
Modele precum KNN, Naive Bayes, SVM sau MLP nu sunt potrivite pentru sarcini de segmentare semantică, mai ales în contextul datasetului utilizat, unde celulele pot apărea rotite, de dimensiuni variabile sau deformate. Din acest motiv, am încercat inițial utilizarea unui model CNN clasic, însă performanța obținută a fost limitată. Ulterior, în urma unor cercetări, am descoperit că arhitectura U-Net este considerată standardul de aur pentru segmentarea imaginilor medicale.

La început, am folosit funcția de pierdere cross-entropy și am evaluat performanța modelului folosind metrici precum precision și recall. Totuși, mi s-a atras atenția că acestea nu reflectă suficient de bine calitatea segmentării, în special în sarcini pixel-wise. Astfel, am trecut la metrici specifice segmentării: coeficientul Dice și distanța Hausdorff.

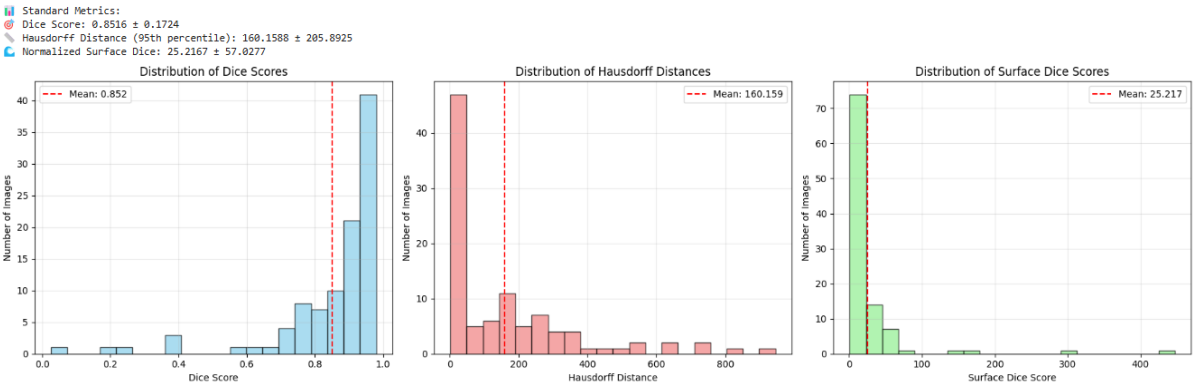
Pentru fine-tuning, am ajustat numărul de filtre din straturile de encoding și decoding. Am observat că arhitecturile mai mici (cu mai puține filtre) și cele care folosesc imagini grayscale oferă rezultate mai bune, probabil datorită reducerii complexității și a overfitting-ului. Arhitectura inițială era destul de mare, cu filtre 64–128–256 și bottleneck 512, și a fost antrenată cu cross-validation. Deși generaliza relativ bine, avea performanță slabă pe tipurile de celule 1, 3 și 4.

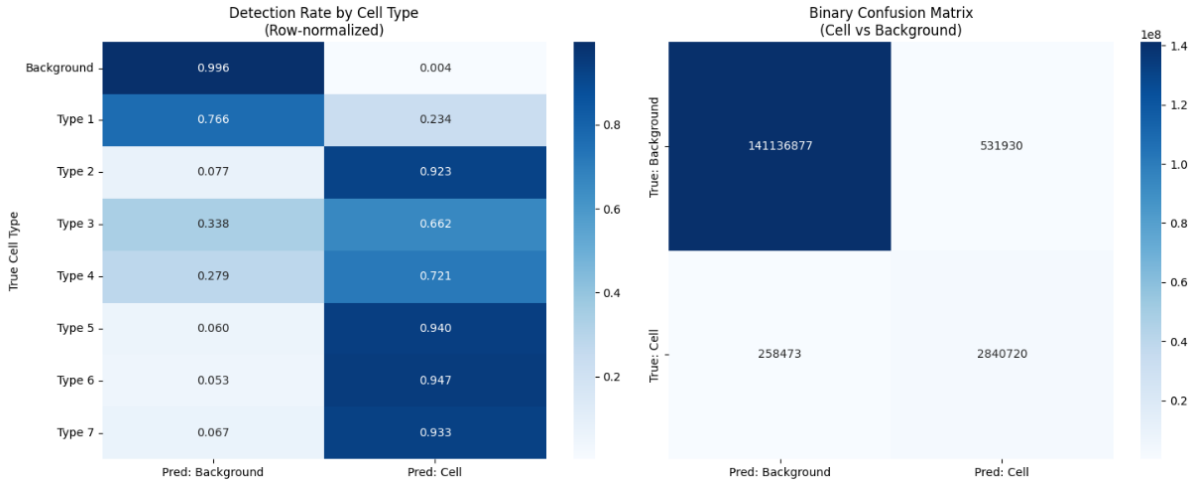






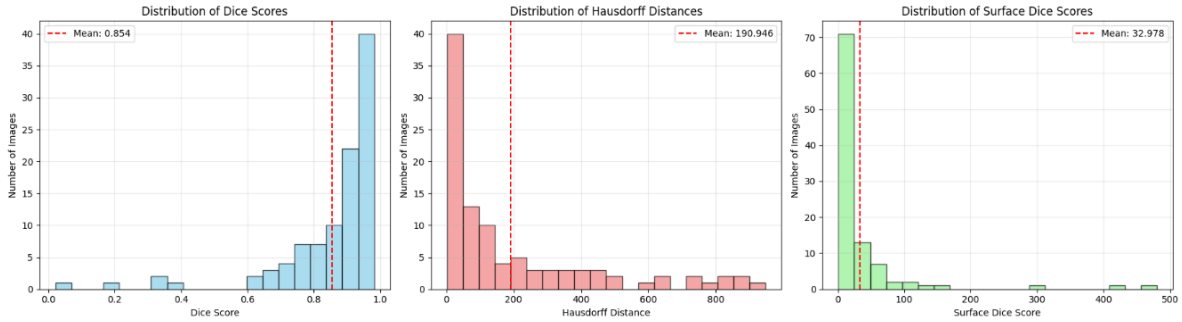
Ulterior, am păstrat arhitectura U-Net, dar am început să experimentez cu funcțiile de pierdere. Într-o primă fază, am înlocuit cross-entropy cu Dice loss. Rezultatele s-au îmbunătățit, în special pe celulele cu contururi clare.

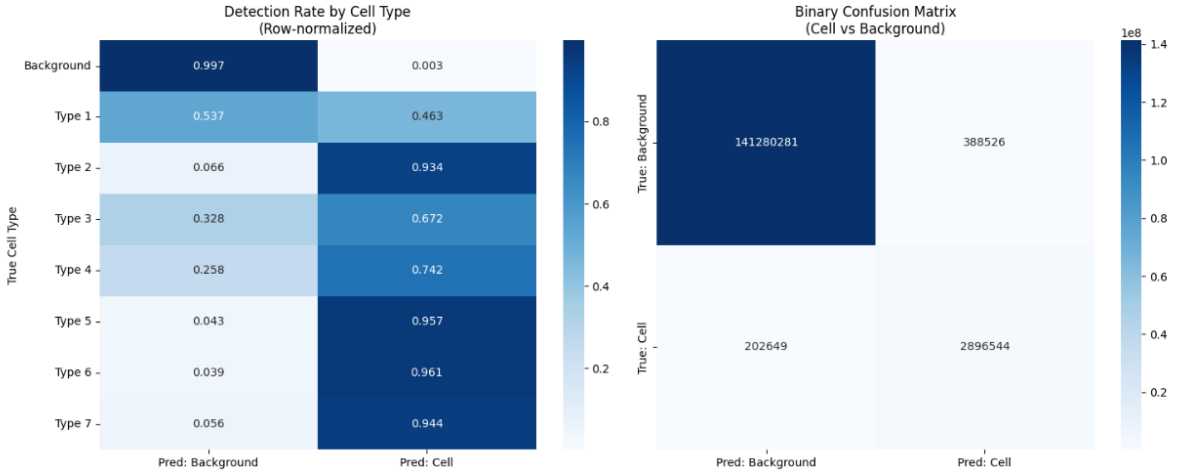




Cel mai bun model intermediar a folosit Dice loss ca funcție de pierdere și o arhitectură

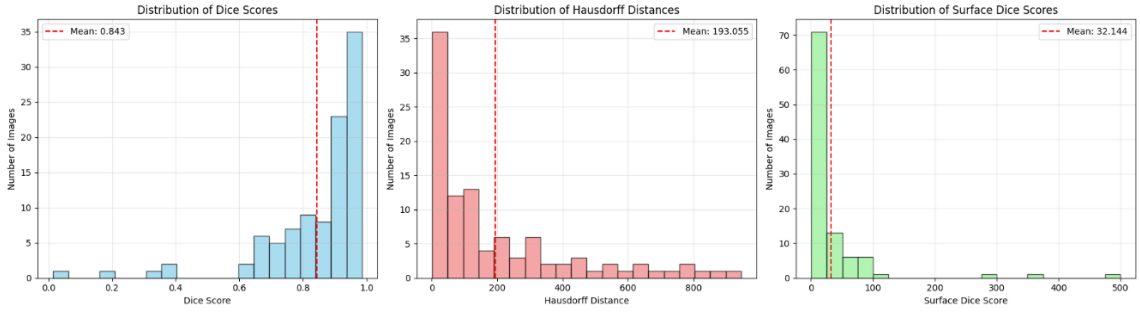
mai redusă (32–64–128 cu bottleneck 256), lucrând pe imagini grayscale. Acesta a obținut un scor de detecție medie de 81.04% pe toate tipurile de celule.

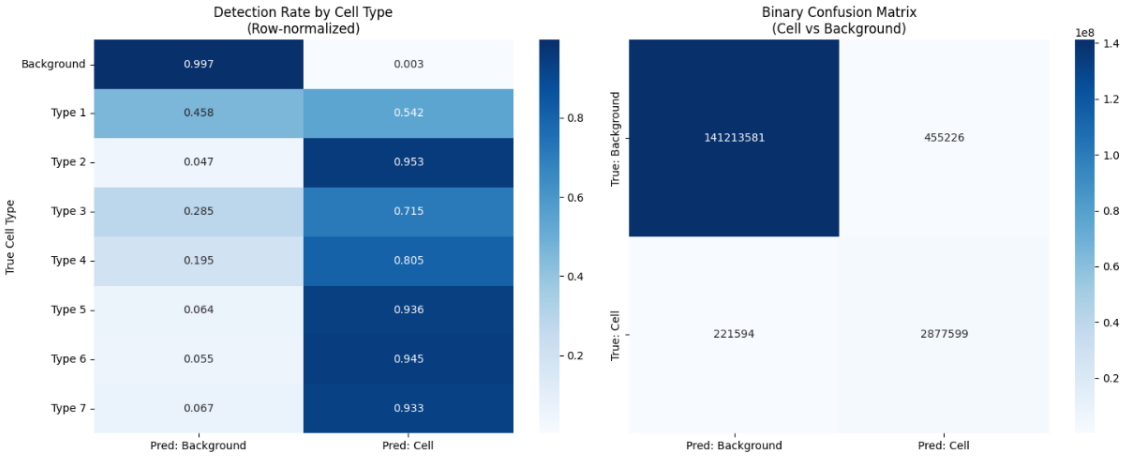




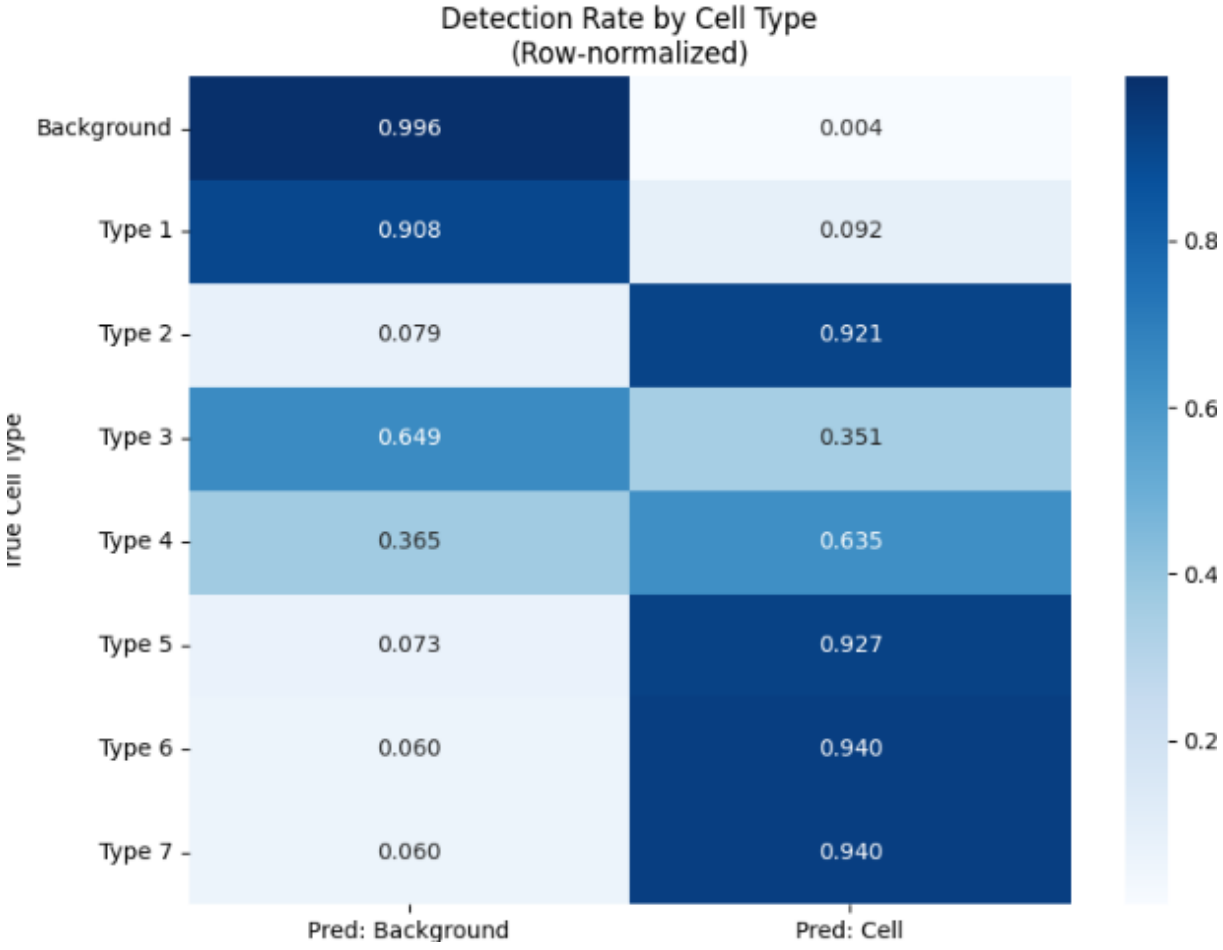
După mai multe experimente, am ajuns la un model optimizat care a atins o rată medie de detecție de 83.27%. Acesta folosește o funcție de pierdere compusă:

Loss = 0.5 * FocalLoss + 0.5 * DiceLoss + 0.0002 * HausdorffDistance Această combinație a reușit să echilibreze foarte bine penalizarea între regiuni greu de detectat și formele complexe ale celulelor.





Am testat și reducerea rezoluției imaginilor (downscaling), însă acest lucru a afectat negativ performanța, în special pe clasele cu puțini pixeli (1, 3, 4), unde acuratețea s-a redus și mai mult.



De asemenea, am aplicat filtrul Scharr în preprocesare (datorită eficienței în detectarea marginilor), dar acesta a crescut semnificativ rata de false positives și a dus la rezultate

mediocre în predicții