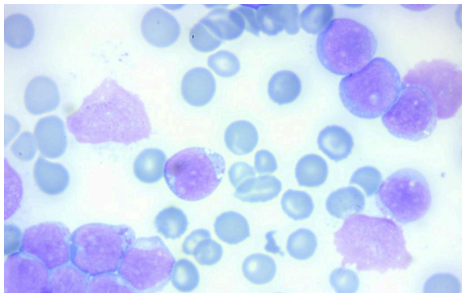
Projecte: Leucèmia

Marc Bardeli Gamez (1637718), Jordi Herrera Valdivia (1638117), Òscar Urenda Moix (1639392)



Abstract—In this project we'll build a prediction script to detect possible cases of leukemia on blood sample images. The script will be able to recognize if a cancer is either benign or one of three kinds malignant.

To achieve that, multiple classifiers and feature extracting methods have been used: CNN, LBP binary, LBP multiclass, HOG, LightGBM. The results are quite good, getting a 97% accuracy on the test.

Keywords— Leukemia, benign, malignant, CNN, LBP, HOG, LightGBM

1 Introducció

El camp de la medicina ha estat un dels principals en ser afavorits per la intel·ligència artificial. Degut a l'existència de models entrenats, és possible un primer anàlisis previ abans que un/a doctor/a faci el seu propi. Tot i que sempre és recomanable que personal qualificat prengui les decisions finals, un cribratge inicial automàtic pot ser d'ajuda.
En aquest projecte en concret es proposa un model d'anàlisi d'imatges de mostres de sang amb presencia de cèl·lules cancerígenes. A través de la seva forma, és possible deduir si es tracten d'un càncer benigne o d'un dels tres possibles tipus de maligne, Pro-B (El càncer està en els primers estats), Early Pre-B (ES la següent fase del càncer, no és molt avançat però ja fa un temps que es desenvolupa en el sistema) I Pre-B(seria un estat avançat del càncer).

2 Estat del Art

Ja hi ha hagut casos on la visió per computador s'ha utilitzat per a la detecció de célules anòmeles [2][3]. Per exemple en la universitat de Osaka han creat una IA basada en xarxes neuronals que es capaç d'agafar mostres de teixits humans i analitzar-les mitjançant visió per computador a partir de un microscopi per observar les cèl·lules i veure si son sanes i anòmales.

També hi ha un altre projecte on ha aconseguit realitzar la detecció de cèl·lules anòmales de sang mitjançant reconeixement d'imatges de cèl·lules.

3 Metodologia

S'ha partit d'una base de dades compresa de 3000 imatges de mostres de sang amb cèl·lules cancerígenes. Per facilitar-ne el reconeixement, les cèl·lules han estat tintades per un major contrast amb el fons.

Aquestes imatges estan organitzades en 4 grups diferents: Benignes, Malignes pro-B, Malignes Pre-B i Malignes early Pre-B.

Només ha fet falta fer un únic tipus de preprocessing de les nostres dades, aquest ha estat el balanceig. Donat que teníem moltes més imatges de cada tipus de malignes que no de benignes, es va decidir a priori agafar 200 imatges de cada tipus de les malignes i totes les benignes que eren unes 500. Amb això vam ajuntar els diferents tipus d'imatges de cèl·lules malignes en un sol grup i vam aplicar el LBP.

Posteriorment, es va fer un altre preprocessing i es va fer per tenir cada grup de cèl·lules per separat, aquí es va agafar el mateix nombre d'imatges de cada grup que les que tenia el de benignes, un cop teníem les dades es va aplicar un altre cop el LBP i seguidament el HOG, per acabar agafant les features del LBP amb totes les classes també s'ha implementat un LightGBM per veure quin mètode era millor i donava resultats més satisfactoris.

Donat que en les execucions d'aquests mètodes esmentats, el temps que l'ordinador es passava treballant era massa, s'ha aplicat un mètode per poder guardar les features d'una execució a memòria i tenir-les sempre per evitar que les hagi de buscar a cada execució. Carregant les features guardades abans de passar per cada mètode, ha disminuït el temps considerablement.

L'últim model que mètode que es va decidir implementar ha estat el CNN donat a l'alta complexitat que té el fet d'analitzar les imatges i saber classificar-les es va optar per una xarxa neuronal capaç d'anar aprenent a cada execució, manualment llavors es va buscar quantes èpoques s'havien de fer per obtenir un resultat òptim, primer es va fer amb les imatges que teníem i seguidament es va tornar a fer un CNN però aquesta vegada havent fet un data augmentation considerable per entrenar-lo

millor. Per tal de poder implementar aquest algorisme hem hagut de transformar les dades a NumPy per poder treballar amb elles de manera més eficient. Per tal de facilitar la detecció de les cèl·lules cancerígenes vam decidir aplicar una binarització a les imatges.

4 Experiments, Resultats i Anàlisi

Primerament, donat que ja teníem les dades preparades per aplicar els mètodes es va fer el LBP binari, és a dir, tenint només si les cèl·lules eren benignes o malignes, per aquest cas es van utilitzar els models: SVM i random forest per a mesurar resultats. Amb svm els resultats no van ser gaire bons, tenint d'accuracy entre un 60-65%, veient això vam pensar en provar un enfocament diferent, aquí va sorgir el random forest, utilitzant-lo vam arribar fins a un accuracy d'entre 87-91% millorant bastant el que teníem.

```
Test Accuracy with Random Forest: 0.8878923766816144
              precision    recall  f1-score   support

         ben       0.88      0.84      0.86        91
         mal       0.89      0.92      0.91       132
```

Figura 1: Resultats de l'execució de Random Forest

Per un anàlisi correcte i visual de les dades es va mirar la matriu de confusió on vam poder veure que l'error de predicció entre les dues classes era molt similar.

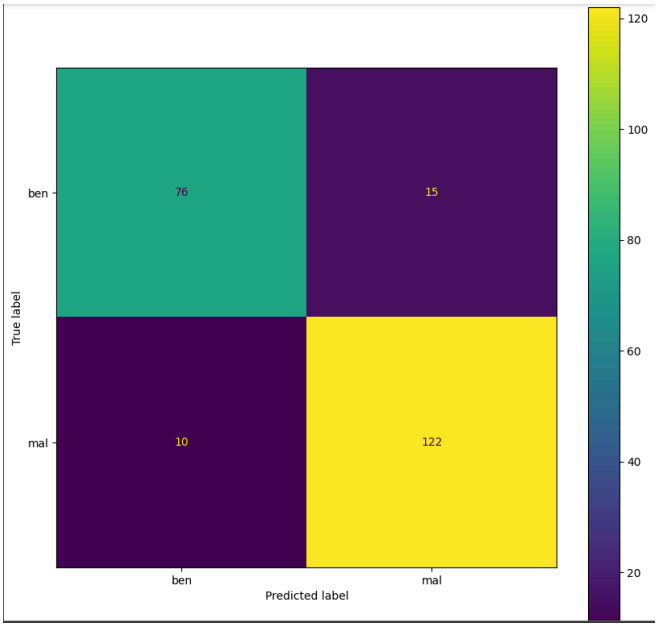


Figura 2: Matriu de confusió dels resultats de Random Forest amb LBP binari

Amb això fet es va decidir continuar explorant resultats, però amb totes les classes separades, per tant, es rellegeixen imatges i se n'afegeixen a cada classe per igualar el nombre d'imatges que té el grup de benignes.

Tenint els grups ja fets es procedeix a tornar a aplicar el LBP amb els mateixos models, dona uns resultats nefastos d'entre 25-30% d'accuracy en el test, cal mencionar que en diferents execucions havia donat resultats bons, però en l'esmentada no, tot i això cal mencionar-la, ja que les features utilitzades aquí també s'utilitzen en el LightGBM amb resultats molt millors.

Seguidament, un altre cop amb random forest i svm es fa un HOG i, per tant, es busquen característiques de forma diferent a fins com s'havia fet fins ara. El resultat de fer el HOG es tradueix en uns nombres pitjors al LBP binari tant en svm com en el random forest, però millors que en el LBP amb tots els grups, els resultats han estat d'entre 50-60% amb què ens fa pensar que és un gran overfitting, ja que el train té un 100% d'accuracy.

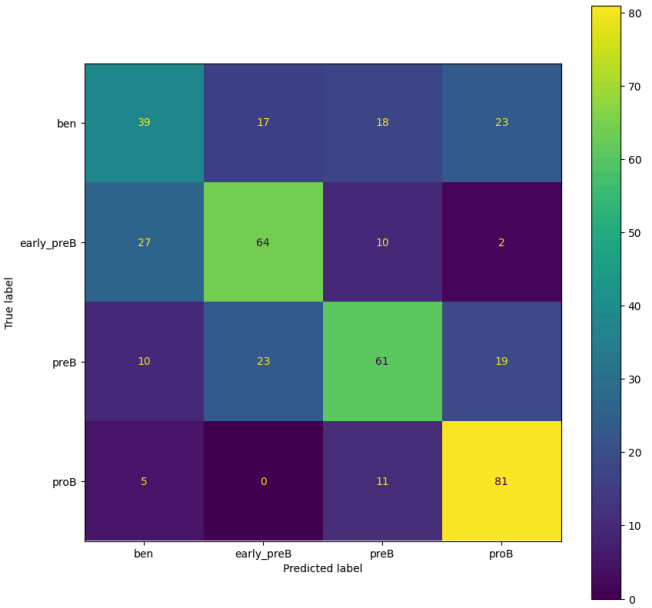


Figura 3: Matriu de confusió dels resultats de Random forest amb HOG multiclasse.

Tenint guardades totes les features es podia fer servir qualsevol, ens vam decantar per les del LBP amb totes les classes. Per fer servir el LightGBM primer s'ha de fer un label encoder de les nostres etiquetes, amb això i posat els paràmetres desitjats per optimitzar-ho al nostre gust ja és va executar. Els resultats van ser satisfactoris, va millorar una mica respecte al que teníem, però amb un 92% d'accuracy encara teníem marge de millora.

Amb tots aquests resultats, veiem que ens estàvem apropant al que buscàvem, però encara ens faltava una mica, al final del dia en una situació mèdica no vols tenir cap dubte de si una cèl·lula és cancerígena o no, per tant, necessitàvem apropar-nos

el màxim possible al 100% d'accuracy. Així que vam pensar que hauríem d'implementar un model més complex que pugui aconseguir millors resultats.

Per fer aquest model més complex vam decidir implementar un algorisme CNN.

Per tal de poder implementar aquest algorisme hem hagut de transformar les dades a NumPy per poder treballar amb elles de manera més eficient. Per tal de facilitar la detecció de les cèl·lules cancerígenes vam decidir aplicar una binarització a les imatges.

Un cop implementat l'algorisme vam decidir fer una prova per trobar les èpoques òptimes que hauria de realitzar la xarxa neuronal on vam arribar als següents resultats:

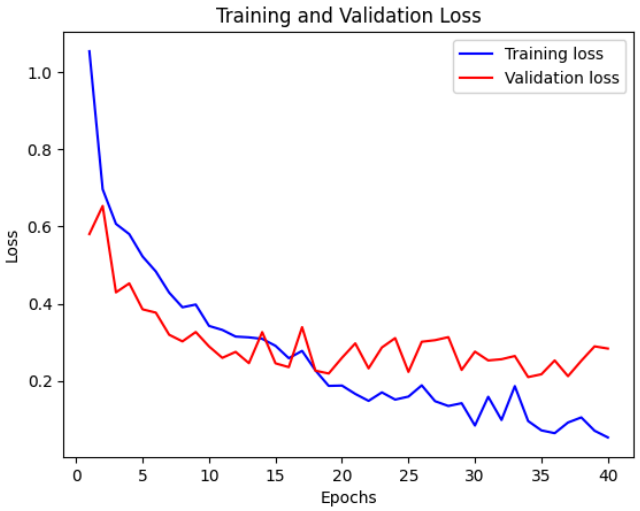


Figura 4: Evolució de les pèrdues segons el transcurs de èpoques.

En veure els resultats obtinguts, on de forma relativament ràpida vam obtenir un 97% d'accuracy vam decidir intentar obtenir el millor resultat possible amb aquest algorisme.

Vam confluir que les imatges de la base de dades podrien arribar a ser insuficients per poder fer que l'algorisme pogués generar un classificador més eficaç, per això vam realitzar una data augmentation.

Aquesta data augmentation voltejara les imatges de forma aleatòria de forma vertical i horitzontal, i també rota les imatges 3 vegades consecutives 90° en direccions aleatòries.
Amb això aconseguim que de cada imatge original n'obtenim 3 de diferents, augmentant el volum de la base de dades i la robustesa de l'aprenentatge de l'algorisme.

Amb tot això, els resultats del CNN van estar els següents:

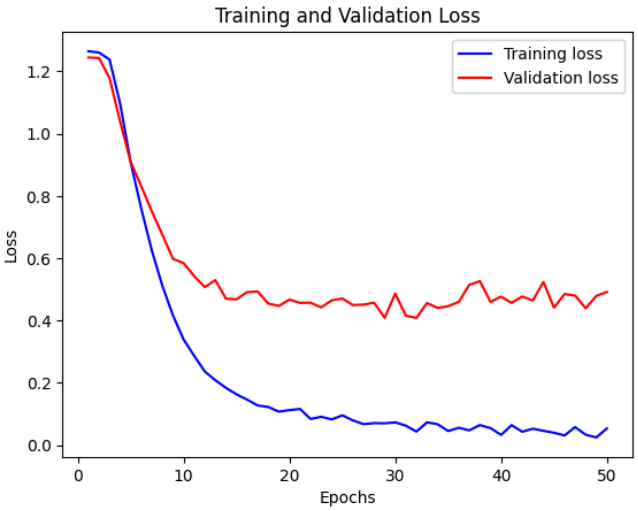


Figura 5: Evolució de les pèrdues segons el transcurs de èpoques amb ús de tècniques de data augmentation

4 Cas d'ús

El programa que hem fet es pot aplicar directament en reconeixement de cèl·lules canceroses, ja que aquest era el nostre objectiu principal, ser capaços de classificar-les. Més enllà d'això també es pot arribar a aplicar per a reconeixement d'altres coses, mentre es tinguin imatges els mètodes implementats han de tenir la capacitat d'identificar característiques en les imatges i classificar-les segons les etiquetes que es tinguin, algun exemple seria mitjançant una prova analítica ordinària, poder detectar si hi ha cèl·lules cancerígenes i el seu tipus.

Si es pogués implementar aquest sistema als centres mèdics es podria fer una detecció més eficaç i poder començar el tractament com més aviat millor.

Aquesta mesura podria potencialment augmentar la taxa de supervivència d'un dels càncers més perillosos existents.

Ara bé, tenint en compte els resultats obtinguts, tot i que l'aplicació de l'algorisme pot resultar altament precís en la detecció, continua essent un algorisme imperfecte, sempre s'haurà de tindre un marge d'error i una supervisió de professionals per a cada cas per evitar possibles errors de predicció.

5 Conclusions

El dataset escollit ha provat ser suficientment ample (Tot i que s'han aplicat mètodes de data augmentation) com per construir models classificadors. La seva forma base, amb les cèl·lules tintades per aïllar-les del fons, ha estat de gran ajuda.

Durant el transcurs del projecte és possible veure una evolució de models més simples a més complexos, basats en xarxes neuronals. Alguns dels classificadors més senzills (HOG) donaven resultats mediocres (55%) mentres que d'altres (LightGBM) obtenien resultats relativament alts (92%), però no era suficient.

Degut a l'objectiu i el camp on s'aplicarien aquests classificadors, era clau obtenir resultats el més acurats possibles i amb una taxa d'error minima. Va estar necessari fer un pas més i avançar a models basats en xarxes neuronals (CNN), on vam trobar millors resultats (97%) que s'aporpen al nostre objectiu.

Encara que la opinió de un professional qualificat és imprescindible a l'hora de diagnosticar aquesta mena de casos, els classificados que hem entrenat poden donar una primera opinió relativament segura.

Bibliografia

[1] Blood Cells Cancer (ALL) dataset - Kaggle

https://www.kaggle.com/datasets/mohammadamireshraghi/blood-cell-cancer-all-4class

[2] Inteligencia artificial para detectar células cancerosas - https://www.consalud.es/

https://www.consalud.es/saludigital/tecnologia-sanitaria/inteligencia-artificial-para-detectar-celulas-cancerosas_58215_102.html

[3] Cancer cells distinguished by artificial intelligence-based system - EurekAlert!

https://www.eurekalert.org/news-releases/502518