Tehnici bazate pe Machine Learning pentru îmbunătățirea depistării cancerului de sân

Elena-Anca PARASCHIV[1,2], Elena OVREIU[2]

[1]Institutul Național de Cercetare-Dezvoltare în Informatică – ICI București

[2]Universitatea Politehnica din București

elena.paraschiv@ici.ro, elena.ovreiu@upb.ro

Rezumat: Cancerul de sân este unul dintre cele mai frecvente tipuri de cancer diagnosticat la femei și ocupă locul al doilea ca și cauză a mortalității provocate de cancer, după cel de plămâni. Atât diagnosticul cât și predicția dezvoltării cancerului de sân sunt realizate, în zilele noastre, folosind diferite tehnici bazate pe metode avansate cum ar fi învățarea automată (Machine Learning). Articolul de față urmărește prezentarea rezultatelor cercetărilor în domeniul Machine Learning aplicate în scopul clasificării datelor medicale. Prin folosirea unui set de algoritmi diferiți s-a urmărit clasificarea bazei de date Breast Cancer Wisconsin pentru diagnostic. Criteriile de selecție ale algoritmilor au fost alese astfel încât să evidențieze performanțele tehnicilor de tip Machine Learning din punctul de vedere al acurateței și al preciziei. Pentru implementare s-au folosit tehnici precum: Mașini cu Vectori Suport (SVM – Support Vector Machines), k-Nearest Neighbor (kNN), Perceptron Multistrat (MLP – Multilayer Perceptron), Arbore Decizional (Decision Tree), Gaussian Naïve Bayes și Random Forest. A fost selectat un set de imagini de diagnostic ale unei tehnici de puncție aspirativă cu ac fin (FNA), din care au fost identificate cele mai reprezentative caracteristici. S-a stabilit că cea mai mare acuratețe a fost obținută în cazul algoritmului Random Forest, în speță 97.90% ceea ce permite conturarea unei perspective de rafinare a clasificării realizate.

Cuvinte cheie: cancer de sân, învățare automată, acuratețe, SVM, kNN, MLP, Arbore Decizional, Random Forest, Gaussian Naïve Bayes.

Machine Learning techniques for an improved breast cancer detection

Abstract: Breast cancer is one the most common types of cancer diagnosed in women and the second leading cause of cancer mortality after lung cancer. The diagnostic and prediction of the cancer development are realized, nowadays, using different techniques based on advanced methods, such as Machine Learning. This article intends to present the research results in the field of Machine Learning applied for the purpose of classifying medical data. Using a set of different algorithms, the aim was to classify the Breast Cancer Wisconsin database for diagnostic. The selection criteria of the algorithms were chosen as to emphasize the performances of Machine Learning techniques in terms of accuracy and precision. For implementation, techniques such as Support Vector Machines (SVM), k-Nearest Neighbor (kNN), Multilayer Perceptron (MLP), Decision Tree, Gaussian Naïve Bayes and Random Forest were used. A set of diagnostic images from a fine needle aspirate technique (FNA) was selected based on which the most representative features were identified. The best accuracy was obtained for the Random Forest algorithm, in this case 97.90%, which allows outlining a perspective of refining the classification achieved.

Keywords: breast cancer, machine learning, accuracy, SVM, KNN, MLP, Decision Tree, Random Forest, Gaussian Naïve Bayes.

1. Introducere

Cancerul, având denumirea științifică de tumoare malignă, reprezintă un grup de anomalii ce implică proliferarea anormală a celulelor în organism (S. Sharma et al., 2018). Spre deosebire de tumorile maligne, cele benigne nu sunt canceroase și nu prezintă riscul de a se răspândi și la alte organe ale corpului. Cancerul de sân reprezintă una dintre principalele cauze ale mortalității la femei, în întreaga lume.

Deși dezvoltarea și îmbunătățirea metodelor de depistare a tumorilor canceroase au evoluat în ultimii ani, cancerul de sân cauzează încă un număr semnificativ de decese din cauza nedepistării la timp a acestuia. Pentru punerea unui diagnostic corect și rapid, există câteva tehnici și metode de predicție, printre care se numără și cele de învățare automată.

https://doi.org/10.33436/v30i2y202006

Machine Learning reprezintă un domeniu al inteligenței artificiale ce furnizează sistemelor abilitatea de a „învăța“ automat și de a căuta anumite modele pe baza datelor analizate, luându-se în final cele mai bune decizii. Scopul principal este acela de a permite calculatorului să învețe singur, fără intervenția sau asistența omului, ajustându-și acțiunile în concordanță cu datele primite (Garry et al., 2018). Învățarea automată utilizează o varietate de algoritmi ce preiau datele și prezic anumite rezultate. Algoritmii se diferențiază în funcție de datele primite: dacă datele sunt etichetate, vom avea algoritmi de învățare supervizată, iar în caz contrar, de învățare nesupervizată.

Inteligența artificială reprezintă una dintre tehnologiile care încearcă să revoluționeze domeniul de sănătate, iar algoritmii de învățare automată pot reuși în curând să detecteze cancerul de sân mult mai bine decât medicii anatomopatologi. În acest articol, se propun mai multe metode de clasificare a setului de date Breast Cancer Wisconsin pentru diagnostic, pe baza unor algoritmi de învățare automată și se determină rezultatele clasificărilor acestora pentru a selecta, în final, cel mai bun algoritm de predicție și detecție a cancerului de sân.

2. Materiale și metode

În acest articol am implementat o serie de algoritmi de învățare automată folosind o bază de date publică numită Breast Cancer Wisconsin (Diagnostic). Acest set de date este compus din 32 de coloane și 569 de linii într-un fișier de tip CSV. Pentru vizualizarea, analiza și procesarea datelor s-a folosit limbajul de programare Python prin intermediul platformei Cloud, furnizată de Google, numită Google Colaboratory. Această platformă utilizează GPU-uri și TPU-uri gratuite, lucru care ușurează crearea de modele destinate învățării automate și, de asemena, lucrează cu Google Drive pentru importarea mai rapidă a fișierelor. Librăriile principale folosite pentru dezvoltarea algoritmilor și vizualizarea rezultatelor au fost: NumPy (pentru algebra liniară), Pandas (pentru deschiderea și procesarea datelor din fișierul CSV), Scikit-learn (pentru implementarea algoritmilor de clasificare și afișarea rezultatelor), Matplotlib (pentru vizualizarea graficelor) și Seaborn (pentru vizualizarea matricei de corelație).

2.1. Descrierea bazei de date Breast Cancer Wisconsin (Diagnostic)

Setul de date Breast Cancer Wisconsin (Diagnostic) este un set de date public, creat de către o echipă de la Universitatea din Wisconsin, SUA și se obține din arhiva UC Irvine Machine Learning Repository. Conține 32 de atribute calculate pentru 569 de instanțe de tumori de sân preluate prin intermediul unei tehnici de puncție aspirativă cu ac fin (FNA). Se prezintă caracteristicile nucleelor celulare din imaginile preluate, iar atributele setului de date sunt calculate pe baza acestora și includ media, eroarea standard și cea mai „gravă“ valoare (media celor mai mari trei valori) a 10 caracteristici ale nucleelor celulare.

Aceste caracteristici sunt:

- Raza (media distanțelor de la centru la punctele de pe perimetru);
- Textura (valorile gray-scale);
- Perimetrul (dimensiunea nucleului tumorii);
- Aria;
- Netezirea (variația locală în lungimea razei);
- Compactitatea ($perimetru^2 / arie - 1$);
- Concavitatea;
- Porțiunile concave;
- Simetria;
- Dimensiunea fractală.

2.2. Importarea și explorarea datelor în limbajul de programare Python

După importarea datelor, prima coloană a tabelului este reprezentată de ID, a doua semnifică diagnosticul pus: M (malign) sau B (benign), iar ultima coloană este NaN, nu prezintă nicio informație.

Pentru deschiderea fișierului s-a folosit funcția *pandas.read_csv* și se afișează în Figura 1 primele 5 linii ale setului de date.

```
Data = pd.read_csv("drive/My Drive/data.csv")#importarea datelor
data.head() #afisarea primelor 5 linii din tabel
```

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean	fractal_dimension_mean
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	0.2419	0.07871
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	0.2069	0.05999
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	0.2597	0.09744
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	0.1809	0.05883

radius_se	texture_se	perimeter_se	area_se	smoothness_se	compactness_se	concavity_se	concave points_se	symmetry_se	fractal_dimension_se
1.0950	0.9053	8.589	153.40	0.006399	0.04904	0.05373	0.01587	0.03003	0.006193
0.5435	0.7339	3.398	74.08	0.005225	0.01308	0.01860	0.01340	0.01389	0.003532
0.7456	0.7869	4.585	94.03	0.006150	0.04006	0.03832	0.02058	0.02250	0.004571
0.4956	1.1560	3.445	27.23	0.009110	0.07458	0.05661	0.01867	0.05963	0.009208
0.7572	0.7813	5.438	94.44	0.011490	0.02461	0.05688	0.01885	0.01756	0.005115

radius_worst	texture_worst	perimeter_worst	area_worst	smoothness_worst	compactness_worst	concavity_worst	concave points_worst	symmetry_worst	fractal_dimension_worst	Unnamed: 32
25.38	17.33	184.60	2019.0	0.1622	0.6656	0.7119	0.2654	0.4601	0.11890	NaN
24.99	23.41	158.80	1956.0	0.1238	0.1866	0.2416	0.1860	0.2750	0.08902	NaN
23.57	25.53	152.50	1709.0	0.1444	0.4245	0.4504	0.2430	0.3613	0.08758	NaN
14.91	26.50	98.87	567.7	0.2098	0.8663	0.6869	0.2575	0.6638	0.17300	NaN
22.54	16.67	152.20	1575.0	0.1374	0.2050	0.4000	0.1625	0.2364	0.07678	NaN

Figura 1. Primele 5 linii din setul de date

După cum se poate observa în Figura 1, tabelul conține 33 de coloane, iar pentru o prelucrare mai bună a datelor se șterge prima coloană ce reprezintă ID-ul fiecărui pacient și coloana Unnamed care nu conține nimic.

Setul de date este constituit din 569 de instanțe dintre care un număr de 357 este reprezentat de clasa **Benign**, iar un număr de 212 de clasa **Malign**, după cum se poate observa în Figura 2 (63% benign și 37% malign).

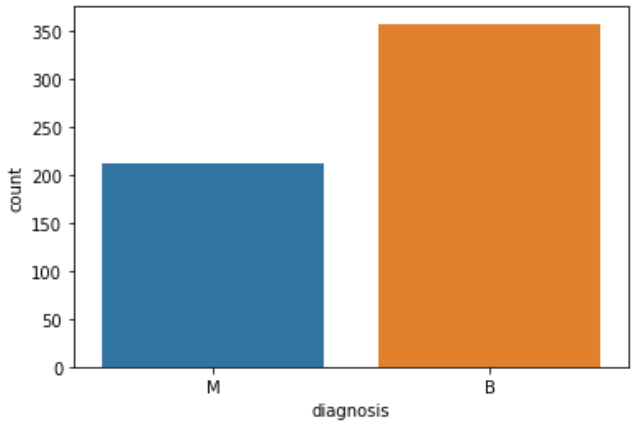

Figura 2. Distribuția claselor setului de date

Pentru o analiză corectă, vom determina corelația dintre diferitele instanțe ale setului de date pentru a desemna apropierea dintre ele și a stabili o relație liniară între acestea.

Aceasta se obține cu ajutorul matricei de corelație care se poate vizualiza în Figura 3.

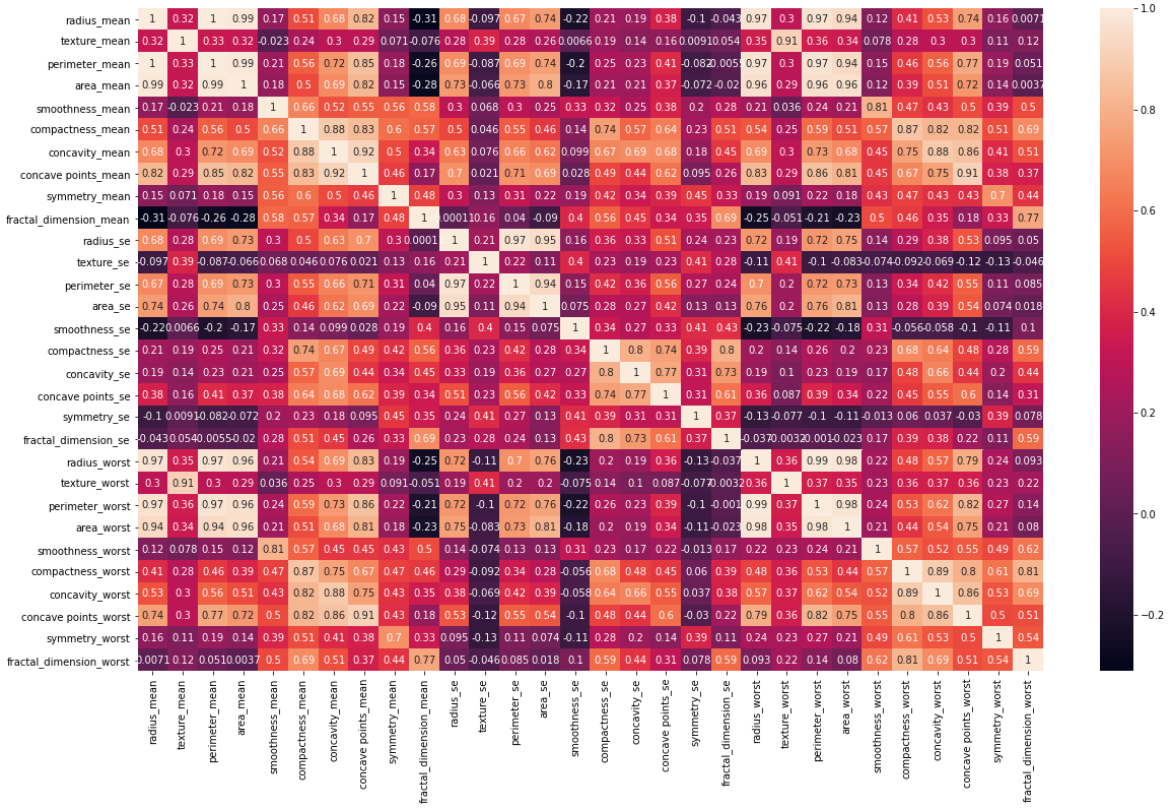

Figura 3. Matricea de Corelație

Se determină matricea de corelație folosind metoda Pearson, valorile prezente cuprinse în intervalul (-1, 1) sugerând corelația dintre atribute. O valoare mai mare de 0.8 indică o corelație puternică între cele două atribute, deci existența unei legi de variație între acestea, iar o valoare mai mică de 0.3 determină lipsa unei relații între respectivele variabile. Astfel, se observă o corelație puternică între radius_mean și perimeter_mean, dar și între area_mean și perimeter_mean.

2.3. Metricile de clasificare

Metricile de clasificare (Dhahri et al., 2019) constau într-o listă de parametri folosiți pentru a evalua performanțele clasificatorului. Pentru calculul acestora se dau următorii parametri:

- TP = True Positive – un exemplar care aparține clasei a fost recunoscut ca aparținând clasei;
- TN = True Negative – un exemplar care nu aparține clasei nu a fost recunoscut ca aparținând clasei;
- FP = False Positive – un exemplar care aparține clasei nu a fost recunoscut ca aparținând clasei;
- FN = False Negative – un exemplar care nu aparține clasei a fost recunoscut ca aparținând clasei.

Sensibilitatea (Recuperarea): reprezintă ponderea dintre exemplarele reale pozitive recunoscute în mod corespunzător de către clasificator ca fiind pozitive.

$$\text{Sensibilitatea} = \frac{TP}{TP + FN} \tag{1}$$

Specificitatea: reprezintă capacitatea unui clasificator de a izola rezultatele negative.

$$\text{Specificitatea} = \frac{TN}{TN + FP} \tag{2}$$

Precizia: proporția identificărilor pozitive făcute corect.

$$\text{Precizia} = \frac{TP}{TP + FP} \tag{3}$$

Acuratețea: reprezintă o măsură a instanțelor corect clasificate.

$$\text{Acuratețea} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

Scorul f1: reprezintă media armonică a valorilor de precizie și de recuperare. Este o măsură a acurateței de test, iar o valoare cât mai apropiată de 1 indică faptul că atât precizia, cât și sensibilitatea sunt foarte bune, ceea ce înseamnă că tehnica de clasificare este performantă.

$$\text{Scorul } f1 = 2 \cdot \frac{1}{\frac{1}{recuperare} \cdot \frac{1}{precizie}} = 2 \cdot \frac{precizie \cdot recuperare}{precizie + recuperare} = \frac{2TP}{2TP + FP + FN} \tag{5}$$

2.4. Algoritmii de clasificare și rezultatele obținute

2.4.1. Clasificatorul SVM (Mașini cu Vectori Suport)

Un algoritm de clasificare supervizat este reprezentat de *Mașini cu Vectori Suport (SVM)*. Această tehnică de clasificare se bazează pe planele de decizie care definesc granițele de decizie.

SVM generează o funcție liniară care ajută în clasificare. Această funcție liniară numită și hiperplan, împarte în mod distinct clasele, folosind puncte specifice și acționează ca o graniță decizională. Punctele care contribuie la rezultatul algoritmului sunt cunoscute drept vectori suport, având în vedere spațiul multidimensional. Planul decizional este format dintr-un set distinct de obiecte cu diferite clase de apartenență. Acest algoritm mapează spațiul de intrare pentru datele non-liniare către spațiul caracteristicilor de dimensiuni superioare, folosind maparea non-liniară, iar membrii clasei sunt împărțiți de non-membri prin construirea unui hiperplan.

Conceptul de „margine“ a fost introdus de SVM pentru fiecare parte a unui hiperplan (hiperplan pozitiv, hiperplan negativ). Hiperplanul marginii maxime izolează practic cele două clase, cum se poate observa în Figura 4. Maximizarea marginilor este considerabil eficientă, deoarece este utilizată pentru determinarea celei mai mari distanțe posibile între hiperplan și probele de pe ambele părți (Ray et al., 2020).

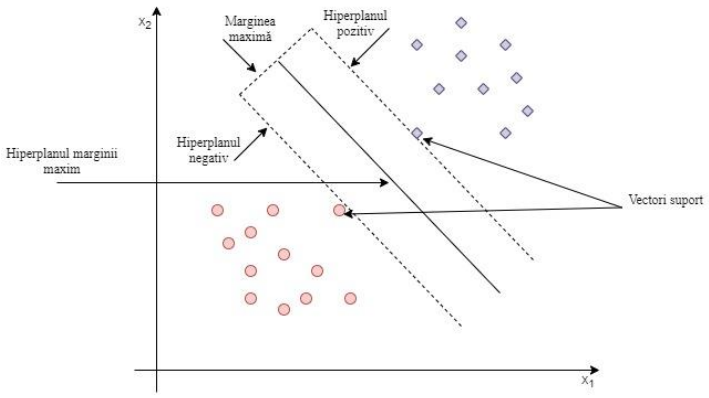

Figura 4. Mașini cu vectori suport

Dacă unul sau mai multe hiperplane sunt disponibile, atunci hiperplanul este ales pe baza marginilor. Marginea reprezintă separarea maximă dintre cel mai apropiat punct de date și hiperplan. Algoritmul SVM este implementat folosind un nucleu (kernel). Predicția se face folosind produsul interior al intrării și fiecare vector suport (care reprezintă coordonatele fiecărui punct de date aflate în apropierea hiperplanului). Nucleele folosite în acest studiu au fost de tip liniar, polinomial și radial.

Având în vedere faptul că lucrăm cu un set de date, pentru implementarea unui clasificator, se împarte setul în: set de antrenare și set de testare. Antrenarea va avea ca scop învățarea de către calculator a caracteristicilor datelor introduse, iar testarea va evalua perfomanțele algoritmului. Această împărțire se face cu ajutorul funcției *train_test_split* din librăria Scikit-learn și se împarte astfel: setul de antrenare de 75% și setul de testare de 25% pentru o antrenare mai bună a clasificatorului.

```
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.25,
 random_state=0)
```

Pentru implementare s-au folosit funcția *SVC* din librăria Scikit-learn și următoarele nuclee: liniar, polinomial și radial. Se obțin următoarele rezultate afișate în Tabelul 1.

Tabel 1. Raportul de clasificare pentru clasificatorul SVM

Nucleul	TP	TN	FP	FN	Acuratețe	Precizie	Sensibilitate	Scorul f1
Liniar	52	85	5	1	95.80%	0.95	0.96	0.96
Polinomial	44	88	2	9	92.30%	0.93	0.90	0.92
Radial	45	89	1	8	93.70%	0.95	0.92	0.93

În tabelul de mai sus sunt evidențiate metricile de clasificare care ajută la evaluarea performanțelor algoritmului. Acuratețea cea mai bună se obține pentru nucleul liniar, de aproape 96%. De asemenea, o sensibilitate mare este de dorit în domeniul medical, astfel încât să nu existe cazuri în care pacientului să i se spună că e sănătos și, de fapt, să fie bolnav.

Matricea de confuzie, care se calculează pe baza parametrilor TP, TN, FP, FN, pentru clasificatorul SVM cu nucleu liniar este prezentată în Figura 5.

	predicted_cancer	predicted_healthy
is_cancer	52	1
is_healthy	5	85

Figura 5. Matricea de confuzie pentru clasificatorul SVM (cu nucleu liniar)

Pe baza matricei de confuzie se poate determina faptul că acest clasificator a prezis corect pentru 52 de pacienți ca având cancer (TP) și 85 de pacienți ca fiind sănătoși (TN).

De asemenea, 5 pacienți au fost clasificați greșit ca fiind bolnavi (FP), ei fiind sănătoși și doar 1 pacient clasificat greșit ca fiind sănătos, el fiind, de fapt, bolnav (FN).

Așadar, se obțin performanțe foarte bune ale clasificatorului SVM, în special pentru nucleul liniar.

2.4.2. Clasificatorul k-Nearest Neighbor („Cei mai apropiați k vecini“)

Un „cel mai apropiat vecin“ k poate fi definit ca algoritmul utilizat pentru a determina în ce clasă se încadrează un set de date pe baza celorlalte seturi de date prezente în jurul său. Această tehnică este o abordare a învățării supervizate folosită pentru regresie și clasificare. Pentru a procesa un nou punct, clasificatorul kNN adună toate punctele apropiate de acesta. Atributele care au un grad mare de variație sunt factori cheie în determinarea distanței.

Fiind dați N vectori de antrenare în Figura 6, algoritmul kNN identifică cei mai apropiați k vecini, indiferent de etichete.

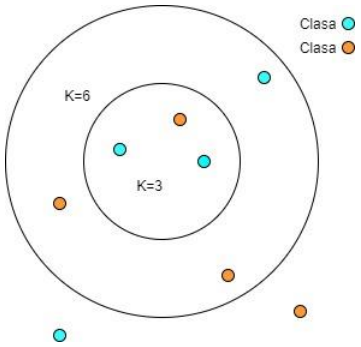

Figura 6. Ilustrație kNN

Clasificatorul kNN ajută la prezicerea proprietăților datelor. Să presupunem că A și B sunt două clase prezente în setul nostru de date. Un nou punct C este adăugat în setul de date și trebuie clasificat pentru a se încadra într-una dintre clase. Mai mult, C este clasificat în funcție de cei k vecini ai săi care cuprind elemente de date majoritare (Kurnianingsih et al., 2016). Măsurarea distanțelor, cum ar fi distanța euclidiană, este utilizată pentru a determina care dintre k vecinii prezenți în setul de date, care urmează să fie studiat, sunt mai comparabili cu elementul care este inclus ulterior. Distanța euclidiană este determinată de rădăcina pătrată a sumei diferențelor pătrate între o poziție nouă (D_1), cu coordonate (x_1, y_1) și o poziție existentă (D_2), cu coordonate (x_2, y_2), luând în considerare toate atributele de intrare pentru compararea distanței dintre elementul de date specific și alte elemente de date.

$$\text{distanța euclidiană dintre } D_1 \text{ și } D_2 = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{6}$$

Pașii algoritmului:

- Se selectează k – numărul de vecini;
- Se calculează vecinii cei mai apropiați de C în funcție de distanța euclidiană;
- Se calculează numărul de puncte de date din fiecare clasă prezentă în cei mai apropiați k vecini;
- Se atribuie punctul C clasei cu numărul maxim de vecini (Amrane et al., 2018).

Pentru implementarea clasificatorului kNN pe setul de date Breast Cancer Wisconsin, se folosește funcția `KNeighborsClassifier`, alegând un k în intervalul (1, 50).

În graficul din Figura 7 se poate observa evoluția acurateței în funcție de k-ul ales din intervalul (1,50). Ceea ce ne interesează este acuratețea obținută pe setul de test, aceasta având valoarea cea mai mare la un k mic și scăzând odată cu creșterea valorii lui k.

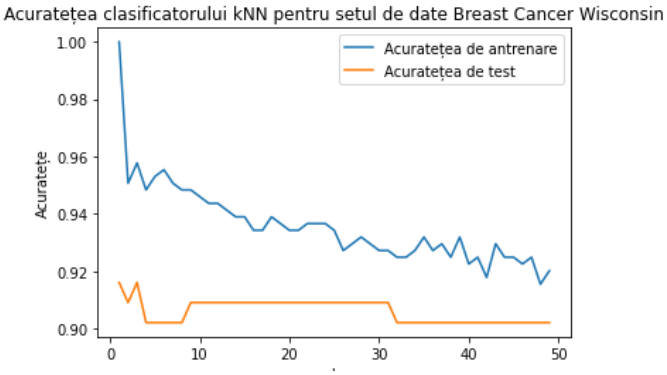

Figura 7. Grafic acuratețe kNN

Pentru exemplificare am ales, pentru k, valorile de 3, 5, 15, 30 și 50 și se obțin următoarele rezultate afișate în Tabelul 2.

Tabel 2. Raportul de clasificare pentru clasificatorul kNN

Valoare K	TP	TN	FP	FN	Acurateţe	Precizie	Sensibilitate	Scorul F1
3	88	43	10	2	91.60%	0.93	0.89	0.906
5	89	40	13	1	90.20%	0.92	0.87	0.889
15	89	41	12	1	90.90%	0.93	0.88	0.897
30	89	41	12	1	90.90%	0.93	0.88	0.897
50	89	40	13	1	90.20%	0.92	0.87	0.889

Având în vedere graficul din Figura 7 și Tabelul 2, se observă că cele mai bune rezultate se obțin pentru un k destul de mic. În cazul de față, pentru k=3 s-a obținut cea mai mare valoare a acuratateței, și anume, de 91.608%.

Matricea de confuzie a clasificatorului kNN pentru k=3, este reprezentată în Figura 8.

	predicted_cancer	predicted_healthy
is_cancer	88	2
is_healthy	10	43

Figura 8. Matricea de confuzie pentru clasificatorul kNN (k=3)

Matricea de confuzie din Figura 8 determină faptul că acest clasificator a prezis corect pentru 88 de pacienți ca având cancer (TP) și 43 de pacienți ca fiind sănătoși (TN). Dar 10 pacienți au fost clasificați greșit ca fiind bolnavi (FP), ei fiind sănătoși și 2 pacienți clasificați greșit ca fiind sănătoși, ei fiind, de fapt, bolnavi (FN).

2.4.3. Clasificatrorul Perceptron Multistrat (MLP)

Perceptronul Multistrat (Multilayer Perceptron – MLP) face parte din categoria rețelelor neuronale supervizate (Vrejoiu, 2019). Acestea sunt definite de existența unui set de antrenare etichetat, în care etichetele asociate vectorilor reprezintă setul de ieșiri ideale ale rețelei atunci când la intrare se aplică vectorul respectiv (Ghosh et al., 2014).

În figura de mai jos (Figura 9) se poate observa structura perceptronului multistrat. Primul strat este denumit strat de intrare – neuronii sunt virtuali astfel încât ei nu realizează prelucrarea semnalului, ci au doar rolul de multiplexor. Stratul următor este cel intermediar sau ascuns, unde se detectează caracteristicile setului de date, iar cel de-al treilea nivel conține neuronii „perceptroni“ sau „clasificatori“ care combină caracteristicile date de stratul intermediar pentru a lua deciziile de recunoaștere a formelor. Prin urmare, prelucrarea semnalului se realizează doar în stratul intermediar și cel de ieșire.

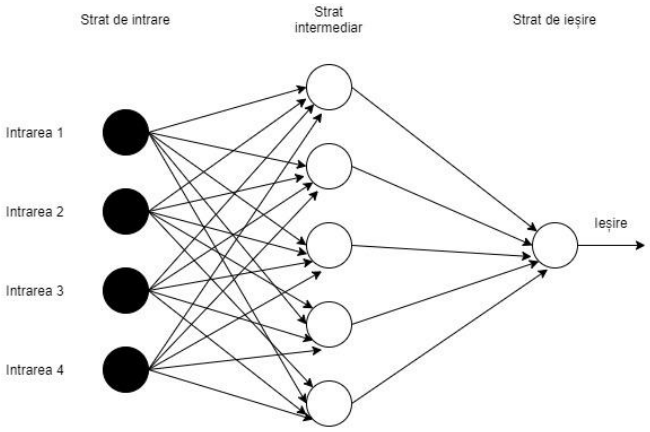

Figura 9. Structura Perceptronului Multistrat

Clasficatorul MLP a fost implementat cu ajutorul funcției MLP Classifier din librăria Scikit-learn pentru 3 funcții de activare diferite, iar rezultatele clasificării sunt afișate în Tabelul 3.

Tabel 3. Raportul de clasificare pentru clasificatorul MLP

Funcția de activare pentru MLP	TP	TN	FP	FN	Acuratețe	Precizie	Sensibilitate	Scorul f1
ReLu (Rectified Linear Unit Function)	47	78	12	6	87.41%	0.86	0.88	0.87
Logistic (Sigmoid function)	52	87	3	1	97.20%	0.97	0.97	0.97
Tanh (Hyperbolic tan function)	51	83	7	2	93.70%	0.93	0.94	0.94

După cum se poate observa în Tabelul 3, cea mai mare acuratețe a clasificatorului MLP se obține pentru funcția de activare *logistic*, ce are ca funcție matematică de bază, funcția sigmoidă. Acuratețea maximă rezultată este de 97.20%, iar sensibilitatea de 0.96 ceea ce face ca acest algoritm să fie un bun clasificator pentru datele medicale.

	predicted_cancer	predicted_healthy
is_cancer	52	1
is_healthy	3	87

Figura 10. Matricea de confuzie pentru clasificatorul MLP (funcția de activare *logistic*)

Matricea de confuzie pentru clasificatorul MLP, cu funcția de activare *logistic,* se poate observa în Figura 10. 52 de pacienți au fost clasificați corect ca fiind bolnavi (TP) și 87 de pacienți clasificați corect ca fiind sănătoși. Doar 3 pacienți au fost clasificați greșit ca fiind bolnavi (FP) și unul singur clasificat greșit ca fiind sănătos, el fiind, de fapt, bolnav.

2.4.4. Clasificatorul Gaussian Naïve Bayes

Naïve Bayes este o tehnică de clasificare bazată pe teorema lui Bayes (Shaikh et al., 2019). Naïve Bayes consideră fiecare atribut ca fiind autonom; se consideră faptul că prezența unei caracteristici specifice într-o clasă nu este identificată cu prezența unor caracteristici diferite. În loc de predicție, clasificatorul Naïve Bayes decide probabilitatea ca datele să fie într-o anumită clasă. Probabilitățile utilizate de Naïve Bayes sunt probabilitatea de clasă și probabilitățile condiționale. Probabilitatea clasei este probabilitatea fiecărei clase, iar probabilitatea condițională este probabilitatea ca fiecare valoare de intrare este dată fiecărei clase.

$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)} \tag{7}$$

unde c și x sunt evenimente și $P(x) \neq 0$. .

$P(c|x)$ este o probabilitate condițională: probabilitatea ca evenimentul c să se producă dacă x este adevărat;

$P(c|x)$ este tot o probabilitate condițională: probabilitatea ca evenimentul x să se producă dacă c este adevărat;

$P(x)$ și $P(c)$ sunt probabilitățile de observare a evenimentelor x și, respectiv c și sunt numite probabilități de margine.

În cazul clasificatorului Naïve Bayes, se presupune că valorile continue asociate cu fiecare caracteristică sunt distribuite în funcție de un model gaussian. O distribuție gaussiană se numește și distribuție normală. Probabilitatea caracteristicilor este presupusă a fi gaussiană, deci, probabilitatea condițională este dată de:

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right) \quad (8)$$

unde μ_y reprezintă media valorilor lui x_i asociate cu evenimentul y, iar σ_y^2 reprezintă varianța corectată Bessel a valorilor x_i asociate cu y.

Implementarea acestui clasificator se face cu ajutorul funcției `GaussianNB` și se obțin următoarele rezultate afișate în Tabelul 4.

Tabel 4. Rezultatele clasificării pentru clasificatorul Gaussian Naïve Bayes

Clasificator	TP	TN	FP	FN	Acuratețe	Precizie	Sensibilitate	Scorul f1
Gaussian Naïve Bayes	48	86	4	5	93.70%	0.93	0.93	0.932

În cazul clasificatorului Gaussian Naïve Bayes, acuratețea rezultată este de aproape 94%. Este un clasificator destul de performant, dar mai slab comparativ cu cele studiate anterior. Matricea de confuzie pentru acesta se poate observa în Figura 11.

	predicted_cancer	predicted_healthy
is_cancer	48	5
is_healthy	4	86

Figura 11. Matricea de confuzie pentru clasificatorul Gaussian Naïve Bayes

Astfel, pe baza matricei de confuzie, se poate relata faptul că 48 și 86 de pacienți au fost clasificați corect ca fiind bolnavi, respectiv sănătoși, dar 4 pacienți au fost clasificați greșit ca fiind bolnavi și 5 clasificați greșit ca fiind sănătoși, ei fiind bolnavi.

2.4.5. Clasificatorul Decizion Tree (Arbore Decizional)

Arborele Decizional este un algoritm de învățare supervizată utilizat pentru clasificare și regresie. Împarte datele în mod recursiv în vederea obținerii atributelor până în punctul în care apare o condiție de oprire. Ramura Arborelui Decizional reprezintă condiția testului, nodurile de decizie descriu proprietățile, iar nodurile frunzelor reprezintă etichetele clasei, cum este reprezentat în Figura 12. Un dezavantaj al Arborelui Decizional este apariția fenomenului de *overfitting* și atunci când lucrează cu date numerice pierde informații (Nematzadeh et al., 2015).

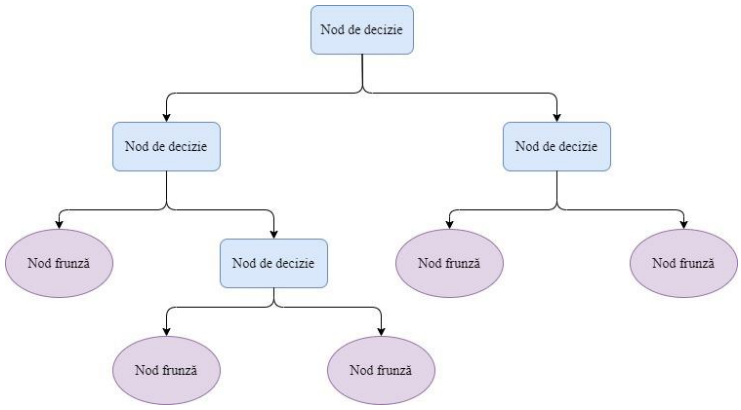

Figura 12. Structura unui Arbore de Decizional

Implementarea acestuia se face cu ajutorul funcției `DecisionTreeClassifier` și se obțin următoarele rezultate afișate în Tabelul 5:

Tabel 5. Rezultatele clasificării pentru clasificatorul Decision Tree

Clasificator	TP	TN	FP	FN	Acuratețe	Precizie	Sensibilitate	Scorul f1
Decision Tree	60	98	10	3	92.39%	0.91	0.93	0.920

Clasificarea folosind algoritmul Decision Tree determină o acuratețe de 92.39%, destul de mică față de algoritmii precedenți. Matricea de confuzie se poate observa în Figura 13.

	predicted_cancer	predicted_healthy
is_cancer	60	3
is_healthy	10	98

Figura 13. Matricea de confuzie pentru clasificatroul Decision Tree

Algoritmul a clasificat corect 60 și 98 de pacienți ca fiind bolnavi, respectiv sănătoși, dar a clasificat 10 pacienți ca fiind bolnavi și 3 ca fiind sănătoși.

2.4.6. Clasificatorul Random Forest (RF)

Random Forest se referă la conceptul de învățare a ansamblurilor care cuprind numeroși algoritmi de învățare automată, compuși pentru a forma un algoritm final optim (Nematzadeh et al., 2015).

Acesta constă dintr-o combinație de numeroși arbori decizionali pentru a grupa o pădure, deoarece modelul se rulează de mai multe ori în comparație cu utilizarea unui singur arbore decizional în care modelul poate fi rulat o singură dată.

În esență, clasificatorul RF urmărește o tactică repetitivă în care un arbore este selectat la întâmplare din subsetul setului de date prezent. Acești arbori decizionali ar putea să nu fie cei mai buni în general, dar împreună pun în aplicare și clasifică destul de bine setul de date.

Există o repetare a etapelor menționate până la obținerea numărului de arbori doriți. După numărarea finală a arborilor, care clasifică observațiile în diferite grupuri, clasificarea cazurilor se bazează pe votul majorității așa cum sunt luate de arborii de decizie.

Un alt motiv pentru care clasificatorul RF este favorizat, se datorează capacității sale de a gestiona valorile marginale ale datelor. În acest caz, o tumoră poate fi clasificată în categorii, în funcție de caracteristicile sale, ea fiind benignă sau malignă.

Implementarea se realizează cu ajutorul funcției `RandomForestClassifier` la care se variază numărul de arbori pentru identificarea celei mai bune clasificări. Rezultatele sunt afișate în Tabelul 6.

Tabel 6. Rezultatele clasificării pentru clasificatorul Random Forest

Numărul de arbori	TP	TN	FP	FN	Acuratețe	Precizie	Sensibilitate	Scorul f1
100	52	88	2	1	97.90%	0.98	0.98	0.98
300	52	87	3	1	97.20%	0.97	0.97	0.97
500	51	87	3	1	96.50%	0.96	0.96	0.96

Conform Tabelului 6, cea mai mare acuratețe de până acum a fost obținută pentru clasificatorul Random Forest cu 100 de arbori decizionali. S-au obținut valori mari atât pentru precizie, cât și pentru sensibilitate, dovedind astfel faptul că acest clasificator este performant în cazul datelor medicale.

În Figura 14 se poate observa și matricea de confuzie specifică acestei clasificări.

	predicted_cancer	predicted_healthy
is_cancer	52	1
is_healthy	2	88

Figura 14. Matricea de confuzie pentru clasificatorul Random Forest

Pe baza figurii de mai sus, se observă că, din totalul de 143 de instanțe specifice setului de testare, 52 de pacienți au fost clasificați corect ca fiind bolnavi, 88 ca fiind sănătoși, iar 2 pacienți au fost clasificați greșit ca fiind bolnavi și doar unul singur clasificat greșit ca fiind sănătos.

Așadar, clasificatorul Random Forest a dovedit cea mai bună performanță dintre clasificatorii prezentați pentru setul de date Breast Cancer Wisconin (Diagnostic).

3. Comparații și discuții

Cei 6 algoritmi de învățare automată au fost implementați pe setul de date Breast Cancer Wisconsin pentru diagnostic folosind ca principale librării Scikit-learn, NumPy și Pandas. Din 569 de instanțe, 143 au fost folosite pentru testare și restul de 426 au fost folosite la antrenarea algoritmilor. În Tabelul 7 se observă cele mai bune performanțe în cazul fiecăruia dintre ele.

Tabel 7. Rezultatele clasificării pentru cei 6 algoritmi

Algoritmul de clasificare	**TP**	**TN**	**FP**	**FN**	**Acuratețe**	**Precizie**	**Sensibilitate**	**Scorul f1**
SVM	52	85	5	1	95.80%	0.95	0.96	0.96
kNN	88	43	10	2	91.60%	0.93	0.89	0.90
MLP	52	87	3	1	97.20%	0.97	0.97	0.97
Gaussian Naïve Bayes	48	86	4	5	93.70%	0.93	0.93	0.93
Decision Tree	60	98	10	3	92.39%	0.91	0.93	0.92
Random Forest	52	88	2	1	97.90%	0.98	0.98	0.98

Așadar, cele mai bune performanțe sunt înregistrate în cazul clasificatorului Random Forest, dar și al clasificatorului MLP (Multilayer Perceptron), ambele obținând procentaje ale acurateței mai mari de 97%.

De asemenea, sensibilitatea, un factor destul de important în cadrul datelor medicale, prezintă rezultate foarte bune atât pentru clasificatorii Random Forest și MLP, cât și pentru clasificatorul SVM.

4. Concluzii

Una dintre cele mai frecvente tumori maligne care poate să apară la femei este cancerul de sân. Astfel, o predicție timpurie în cazul acestei boli poate salva viața multor femei din toată lumea. În acest context, se încearcă implementarea diverselor metode de învățare automată pentru predicția și diagnosticarea din timp a cancerului de sân.

Una dintre metode a fost prezentată în acest articol și a presupus realizarea clasificării setului de date Breast Cancer Wisconin cu ajutorul mai multor tehnici de clasificare automată, supervizată. Algoritmii folosiți au fost SVM, kNN, MLP, Gaussian Naïve Bayes, Decision Tree și Random Forest, iar cele mai bune performanțe au fost obținute de către Random Forest, cu o acuratețe de aproape 98%, dar și pentru clasificatorul MLP cu o acuratețe de 97.02%, dovedind faptul că cei doi sunt cei mai eficienți în cadrul clasificării propuse.

Astfel, a fost demonstrat faptul că tehnicile de învățare automată supervizată sunt performante și pot ajuta și susține un diagnostic mai rapid și eficient al cancerului de sân. În acest context, în perioada următoare se are în vedere diversificarea bazei de date și efectuarea de studii suplimentare mai aprofundate care alături de combinarea tehnicilor prezentate în acest articol pot conduce la rafinarea clasificării deja elaborate.

BIBLIOGRAFIE

1. Amrane, M., Oukid, S., Gagaoua, I. & Ensari, T. (2018). *Breast cancer classification using machine learning*. 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT) (pp. 1-4).
2. Choy, G., Khalilzadeh, O., Michalski, M., Do, S, Samir, A.E., Pianyky, O.S., Geis, J.R., Pandharipande P.V., Brink, J.A. & Dreyer K.J. (2018). *Current Applications and Future Impact of Machine Learning in Radiology*. Radiology, vol. 288,2: 318-328.
3. Dhahri, H., Al Maghayreh, E., Mahmood, A., Elkilani, W. & Nagi, M. (2019). *Automated Breast Cancer Diagnosis Based on Machine Learning Algorithms*. Journal of Healthcare Engineering. 4253641.
4. Ghosh, S., Mondal, S. & Ghosh, B. (2014) *A comparative study of breast cancer detection based on SVM and MLP BPN classifier*. In Proceedings of the 2014 First International Conference on Automation, Control, Energy and Systems (ACES) (pp. 1-4).
5. Kurnianingsih, K., Nugroho, L.E., Widyawan, W., Lazuardi L. & Prabuwono, A.S. (2016). *Emergency alert prediction for elderly based on supervised learning*. In Proceedings of the 2016 1st International Conference on Biomedical Engineering (IBIOMED) (pp. 1-6).
6. Nematzadeh, Z., Ibrahim, R. & Selamat, A. (2015). *Comparative studies on breast cancer classifications with k-fold cross validations using machine learning techniques*. In Proceedings of the 2015 10th Asian Control Conference (ASCC) (pp. 1-6).
7. Ray, A., Chen, M. & Gelogo, Y. (2020). *Performance Comparison of Different Machine Learning Algorithms for Risk Prediction and Diagnosis of Breast Cancer*. In: Fiaidhi J., Bhattacharyya D., Rao N. (eds) Smart Technologies in Data Science and Communication. Lecture Notes in Networks and Systems, vol 105. Springer, Singapore.
8. Shaikh, T.A. & Ali, R. (2019). *Applying Machine Learning Algorithms for Early Diagnosis and Prediction of Breast Cancer Risk*. In: Krishna C., Dutta M., Kumar R. (eds). În Proceedings of 2nd International Conference on Communication, Computing and Networking. Lecture Notes in Networks and Systems, vol 46. Springer, Singapore.
9. Sharma, S., Aggarwal, A. & Choudhury, T. (2018). *Breast Cancer Detection Using Machine Learning Algorithms.* In Proceedings of the 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS) (pp. 114-118).
10. Vrejoiu M. (2019). *Reţele neuronale convoluţionale, Big Data şi Deep Learning în analiza automată de imagini*. Revista Română de Informatică şi Automatică (Romanian Journal of Information Technology and Automatic Control, ISSN 1220-1758, vol. 29(1), pp. 91-114.
11. *** Breast Cancer Wisconsin (Diagnostic) Data Set, 1995: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic) (accesat 2020).

Elena-Anca PARASCHIV activează ca inginer de sistem software în cadrul Serviciului „Aplicații Digitale în Sănătate” al Institutului Național de Cercetare-Dezvoltare în Informatică - ICI București și este studentă la masterul „Sisteme Inteligente și Vedere Artificială” din cadrul Facultății de Electronică, Telecomunicații și Tehnologia Informației, Universitatea Politehnica din București. A absolvit Facultatea de Inginerie Medicală din cadrul Universității Politehnica din București, iar domeniile și subiectele sale de interes pentru activitatea de cercetare cuprind: inteligența artificială cu aplicații în medicină (prelucrare și analiză de imagini și date medicale), telemedicina și dezvoltarea de echipamente pentru asistență medicală.

Elena-Anca PARASCHIV is a Software Engineer in the “Digital Applications in Health” Service at ICI Bucharest and a master’s student in “Intelligent Systems and Computer Vision” at the Faculty of Electronics, Telecommunications and Information Technology, University Politehnica of Bucharest. She graduated from the Faculty of Medical Engineering, University Politehnica of Bucharest and the research fields and topics of interest include artificial intelligence with applications in medicine (processing and analysis of medical images and medical data), telemedicine and the development of healthcare equipment.

Elena OVREIU a absolvit Facultatea de Electronică, Telecomunicaţii și Tehnologia Informației din cadrul Universității Politehnica din Bucureşti în 2009. În paralel a urmat şi cursurile Facultăţii de Relaţii Economice Internaţionale din cadrul ASE, pe care a finalizat-o în Singapore, la la Nanyang Technological University, acolo unde și-a realizat proiectul de diplomă în prelucrare de imagini. A continuat cu un doctorat în Franţa în imagistică computaţională şi apoi s-a perfecţionat în domeniu prin stagii de practică în diverse alte universităţi din Israel, Columbia și China. Din 2013 este lector la Universitatea Politehnica din Bucureşti, antreprenor şi preşedinte al GOMITech, ONG care susţine inovaţia în tehnologie medicală.

Elena OVREIU graduated from the Faculty of Electronics, Telecommunications and Information Technology, University Politehnica of Bucharest in 2009. At the same time, she attended the courses of the Faculty of International Economic Relations from ASE Bucharest. She graduated from Nanyang Technological University in Singapore, where she completed her degree project in image processing. She continued with her PhD in France specializing in computational imaging and then she completed her internship in other universities from Israel, Colombia and China. Since 2013 she has been a lecturer at the University Politehnica of Bucharest, entrepreneur and president of GOMITech, an NGO that supports innovation in medical technology.