cancer la sân: predicția ariei tumorii și clasificare

PROIECT DE DIPLOMĂ

Autor: **Anca-Elena ANDREESCU**

Conducător științific: **Prof. dr. ing. Eva-Henrietta DULF**

|  |  |
| --- | --- |
| DECAN  **Prof. dr. ing. Liviu MICLEA** | Vizat,  DIRECTOR DEPARTAMENT AUTOMATICĂ  **Prof. dr. ing. Honoriu VĂLEAN** |

Autor: **Anca-Elena ANDREESCU**

Cancer la sân: predicția ariei tumorii și clasificare

1. **Enunțul temei:** *Realizarea unui model care poate să ofere o predicție cât mai corectă a ariei tumorii, pentru cancerul la sân. De asemenea, pentru același set de date, s-a dezvoltat un model, pentru a oferi un diagonistic, dacă tumoarea are potențial malign sau benign.*
2. **Conținutul proiectului:** *(enumerarea părților componente) Pagina de prezentare, Declarație privind autenticitatea proiectului, Sinteza proiectului, Cuprins, Titlul capitolului 1, Titlul capitolului 2,… Titlul capitolului n, Bibliografie, Anexe.*
3. **Locul documentării:** *Universitatea Tehnică din Cluj-Napoca*
4. **Data emiterii temei:**
5. **Data predării:**

Semnătura autorului

Semnătura conducătorului științific

**Declaraţie pe proprie răspundere privind**

**autenticitatea proiectului de diplomă**

Subsemnatul(a) **Anca-Elena ANDREESCU**  , legitimat(ă) cu CI/BI seria VX nr. 943559 , CNP 6020317385573 ,

autorul lucrării:

Cancer la sân: predicția ariei tumorii și clasificare

elaborată în vederea susținerii examenului de finalizare a studiilor de licență la **Facultatea de Automatică și Calculatoare**, specializarea **Automatică și Informatică Aplicată,** din cadrul Universității Tehnice din Cluj-Napoca, sesiunea Iulie 2023 a anului universitar 2022-2023, declar pe proprie răspundere, că această lucrare este rezultatul propriei activități intelectuale, pe baza cercetărilor mele și pe baza informațiilor obținute din surse care au fost citate, în textul lucrării, și în bibliografie.

Declar, că această lucrare nu conține porțiuni plagiate, iar sursele bibliografice au fost folosite cu respectarea legislației române și a convențiilor internaționale privind drepturile de autor.

Declar, de asemenea, că această lucrare nu a mai fost prezentată în fața unei alte comisii de examen de licență.

În cazul constatării ulterioare a unor declarații false, voi suporta sancțiunile administrative, respectiv, *anularea examenului de licență*.

Data Prenume NUME

(semnătura)

**SINTEZA**

proiectului de diplomă cu titlul:

Cancer la sân: predicția ariei tumorii si clasificare

Autor: **Anca-Elena ANDREESCU**

Conducător științific: **Titlu. ing. Prenume NUME**

1. Cerințele temei: Realizarea unui model care e capabil să ofere o acuratețe mărită atât în procesul de predicție a ariei tumorii, cât și pentru partea de caracterizare a tumorii: malignă sau benignă.

2. Soluții alese: Utlizarea mediului de dezvoltare Pycharm. Cu ajutorul limbajului de programare Python, s-a antrenat o rețea neuronală artificială (ANN) care rezolvă problema de regresie bazata pe predictia ariei. O altă rețea artificială a fost construită pentru rezolvarea problemei bazate pe împărțirea tumorii in cele doua clase: malignă si benignă

3. Rezultate obținute: Pentru partea de regresie s-a obținut o eroare medie pătratică de de 0.0001 si un coeficient de determinare( R^2) de 0.99. De asemenea, pentru reteau neuronală care se ocupă de împărțirea tumorii în două clase, s-a obținut o acuratețe de 98%.

4. Testări și verificări: Din momentul în care retele neuronale au fost construite, s-au încercat mai multe combinații între parametrii, astfel s-au modificat valori pentru: numărul de straturi ascunse, numărul de neuroni de pe starturi, epoci. De asemenea, setul de date a jucat un rol important în antrenarea rețelei. Verificările s-au pus observa în valoarea erorii mediei pătratice, coeficientul de determinare și în modul în arătau graficele intre datele care se dorea sa fie prezise, și cele prezise.

5. Contribuții personale: Documentarea asupra modului optim de construire a rețelelor neuronale.

6. Surse de documentare:

Semnătura autorului

Semnătura conducătorului științific

Cuprins

[1 Introducere 2](#_Toc167962205)

[1.1 Context general 2](#_Toc167962206)

[1.2 Obiective 3](#_Toc167962207)

[1.3 Specificații 4](#_Toc167962208)

[2 Studiu bibliografic 5](#_Toc167962209)

[3 Analiză, proiectare, implementare 6](#_Toc167962210)

[4 Concluzii 7](#_Toc167962211)

[4.1 Rezultate obținute 7](#_Toc167962212)

[4.2 Direcții de dezvoltare 7](#_Toc167962213)

[5 Reguli de formatare 8](#_Toc167962214)

[5.1 Formatarea paginii 8](#_Toc167962215)

[5.2 Titluri și stiluri 8](#_Toc167962216)

[5.3 Figuri, tabele și ecuații 9](#_Toc167962217)

[5.3.1 Figuri 9](#_Toc167962218)

[5.4 Tabele 9](#_Toc167962219)

[5.5 Ecuații 9](#_Toc167962220)

[5.6 Referințe bibliografice 10](#_Toc167962221)

# Introducere

## Context general

Cancerul de sân a reprezentat dintotdeauna unul dintre principalele motive de deces în răndul femeilor, aproximativ 15% din numărul total. Din acest motiv, importanța găsirii unui diagnostic corect v-a putea conduce la un tratament eficient aplicat pacientelor. [1] Șansele de viață a femeilor pot să fie influențate major de momentul în care este confirmat rezultatul medical, astfel se evidențiază importanța cunoașterii timpurie a diagonisticului.

De-a lungul anilor, aria medicinei s-a dezvoltat foarte mult , astfel reusind să ofere tratamente potrivite si mult mai eficiente pentru pacienti. Totusi, învățarea automată poate să aducă un beneficiu considerabil în clasificarea tumorii, dar si pentru modul în care aria tumorii o să evolueze pe parcursul timpului. [1] O astfel de predicție eficientă asupra particularității tumorii poate să îi ajute si pe medicii oncologi să poată oferii o îngrijire medicală adecvată. Pe parcursul anilor s-a încercat introducerea învățării automate în ajutorul medicinei, iar acest lucru a dus la lucruri inovatoare.

Inteligența artificială (AI) a devenit tot mai prezentă în viața de zi cu zi a oamenilor. Fiecare persoană care deține un telefon se poate numi că a folosit în inteligenta artificială. Fiind un domeniu în continuă dezvoltare și reprezentând o sferă de interes pentru din ce în ce mai mulți programatori, se încearcă răspăndirea ei pe diverse domenii de activitate. Învățarea automată este un subdomeniu al inteligenței artificiale. Cu ajutorul acestei tehnologii se pot crea anumiți algoritmi care pot să învețe cum să ofere o predicție cât mai aproape de adevăr, pe baza unor anumitor seturi de date și relații între aceste date.

O metodă foarte des întălnită pentru învățarea automată este reprezentată de rețelele neuronale artificiale (ANN). Această metoda încearcă se simuleze modul de gândire a oamenilor. Creierul uman este programat să ofere o capacitate mare acumulare a înformațiilor. Acesta este o organizare de aproximativ 10 miliarde de neuroni care comunică între ei cu ajutorul sinapselor. Fiecare celulă de neuron are posibilitatea de a primi, procesa și învăța o informație.[4] Un ANN este format în principiu din trei straturi: stratul de intrare, straturile ascunse și stratul de ieșire. Stratul de intrare este cel care conține informațiile care vor participa la procesul de învățare. Straturi ascunse fac legătura între stratul de intrare și stratul de ieșire. Stratul de ieșire reprezintă de fapt rezultatul la care dorim sa ajungem, în cazul lucrarii prezente dacă tumoarea are o caracteristică canceroasă sau nu, și aria suprafeței tumorii. Ca și în cazul oamenilor, aceste rețele au împărțită informația în două, una pe care învață și o parte pe care testează ce a învățat. De asemnea, rețelele folosesc modul de repetare a informației, trece de mai multe ori prin testul de antrenare ca să ajusteze ce a învățat. Tot acest proces este asemănător cu ce se petrece în creierul unui om în momentul în care învață un lucru nou.

Prin această lucrare, s-au atins două puncte vitale care caraterizează o tumoare: modulul în care o să se dezvolte aria tumorii, bazat pe anumite intrări, respectiv dacă aceasta are o caracteristică malignă sau benignă. Tumoarea reprezintă o acumularea excesivă de celule. Principalul motiv pentru care se formează fiind reprezentat de o acțiunea celulelor de a se divide anormal de mult într-un timp scurt sau din simplu fapt că acestea nu mor într-un timp așteptat. Tumoarea benignă în mod normal nu cauzeaza probleme pacienților, dar chiar dacă au un ritm lent de creștere, pot să ajungă în punctul în care să influențeze funcționarea normală a celorlalte organe din corp uman. Pe de altă parte, tumoarea malignă se caracterizează printr-un ritm rapid si necontrolat de răspândire. Ele reprezintă un real pericol pentru corpul uman, reușind să se răspandească local, dar reușind să ajungă în tot organismul uman prin intermediul sângelui.[2] Aceste aspecte arată importanța cunoașterii tipului tumorii, din acest motiv în această lucrare s-a studiat și încercat găsirea unui algoritm cât mai potrivit pentru aceasta sarcină. De sigur, cunoscând caracteristica tumorii maligne, creștere sporadică, se justifică și dorința medicilor de a cunoaște maniera de evoluție a mărimii tumorii. Această lucrare reflectă dorința de a ajuta și de a ușura modul de gestionare a tratamentului oncologic aplicat oamenilor.

Lucrarea conține capitole în care sunt explicate concis partea de implementare a codului, concluzii, dar și testele realizate asupra algoritmilor obținuți. Setul de date pe care s-au realizat aceste rețele neuronale este obținut de pe UI machine learning, sustrase dintr-o imagine digitalizată a unei mase mamare. Baza de date se numeste Breast Cancer Wisconsin (Diagnostic). Acesta include 569 de instanțe care descriu caracteristicile celulelor. [3]

## Obiective

Lucrarea de față are două obiective principale, să ofere o întelegere mai profundă a modului in care o sa se dezvolte aria tumorii prezente în zona mamară, în plus clasificare a tipul tumorii cu o acuratețe crescută.

Predicția dimensiunii tumorii si clasificarea tipului acesteia, au fost realizate cu ajutorul unui set de date de tip numeric. Acestea conțin informații importante legate de diametru, textură, rază ( distanța medie de la punctele de pe perimetru), compactitate, concavitate (severitatea porțiunilor concave ale conturului), puncte concave (numărul de porțiuni concave ale conturului), simetrie, netezimea ( variația locală a lungimilor razei), diagnostic ( B = benignă, M= malignă), aria tumorii.

Pentru reteaua neuronala care reaizeaza clasificarea tumroii s-a folosit ca si output coloana de „ diagnostic”, iar pentru cea care se ocupă cu pedicția de arie s-a folosit coloana cu numele „arie”.

Toate aceste informații provenite de la mase mamare, au contribuit la realizarea rețelelor neuronale artificale, avand ca și scop final perfecționarea diagnosticului și tratamentului cancerului la sân, furnizând resurse medicale asistate de calculator care au capacitatea de a sprijini medicii în luarea deciziilor clinice. Prin aceste modele, medicii ar avea posibilitatea să beneficieze de o privire de ansamblu asupra modului în care urmează să gestioneze situația în care se află pacientul. Tot odata, aceste modele au ca și scop să elimine eroarea umană, riscul unei diagnosticări eronate.

## Specificații

Domeniul pe care îl vizează lucrarea de față, este cel medical, venind în sprijinul doctorilor de pe secția de oncologie care oferă tratamente pentru femeile care suferă de afecțiuni canceroase în zona sânilor. Cunoașterea tipului tumorii, malignă sau benignă, impactează tipul medicamentației oferit către administrarea persoanelor. Medicul având o viziune pe ansamblu asupra modului în care trebuie abordată problema de tratare, v-a putea să își folosească cunoștiințele medicale pentru a oferi o medicamentație sau o solutie fezabilă pentru a crește șansele de viață a personelor de sex feminin.

Pentru un asemenea domeniu este foarte important să nu se greșeasca diagnosticul, fiind boală care se agravează rapid. De asemenea, este foarte periculoasă din motivul că poate să nu aiba simptome vizibile care să ofere un semnal de alarmă, astfel femeile putând ajunge să descopere destul de tărziu că au o asemenea patologie prezentă în regiunea mamară. Din acest motiv, pentru a putea să fie folosite modele obținute trebuie să aibă o acuratețe suficient de mare și să fie apte să ofere rezultate cât mai apropiate de cele din viața reală. Eroarea de diagnostic trebuie să fie scăzută exponențial, din acest motiv introducerea învățării automate în sfera medicala ar putea minimiza eforturile medicilor, dar ar și micșora tratamente care nu o să aibă rezultate.

Astfel, printr-o implementare avansată, lucrarea ar trebui să furnizeze un mijloc de ajutor care să se faciliteze modul de lucru pentru secțiile destinate tratării afecțiuniilor oncologice. Evaluarea medicală, reușind să se efectueze mult mai optim din punct de vedere al timpului, dar si al tratatemntului, poate să ducă o la creștere de supraviețuire crescută. Calitatea îngrijirii medicale ar trebui să cunoască o îmbunătățire semnificativă pentru impactănd la răndul ei experiența pacienților pe parcursul vindecării.

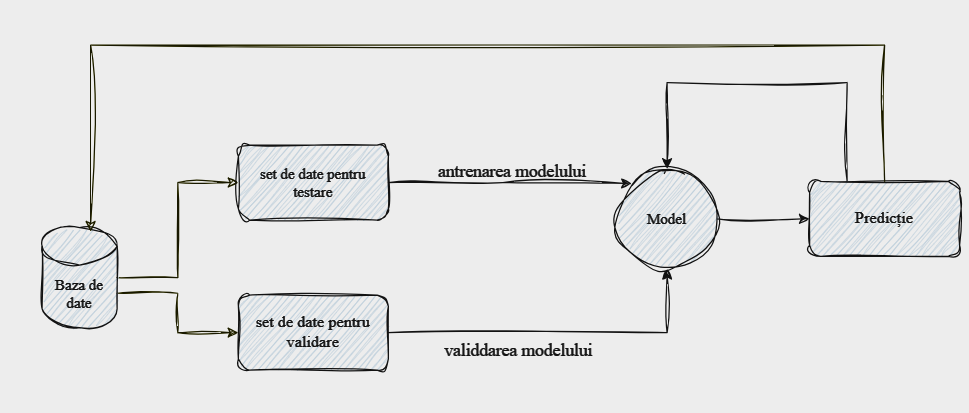
# Studiu bibliografic

Învățarea automată este studiul care se ocupă cu crearea de algoritimi bazați pe realizarea unei sarcini fară a fi programați expliciți pentru acest lucru. Ea a fost explorată cu scopul de a ușura viața oamenilor și de a oferi soluții mai optime la probleme avansate. Principala sursă care oferă posibilitatea învățării automate să funcționeze atât de bine o reprezintă seturile de date voluminoase. Având în vedere cantitatea mărită de seturi de date care există în ziua de astăzi, învățarea automată a reușit să se impună ca fiind o alternativă fezabilă în soluționarea eficinetă si simplificată a problemelor bazate pe date.[5]

Setul de date folosit în lucrarea prezentată este numit Breast Cancer Wisconsin (Diagnostic), furnizând un set amplu de date numerice care au servit la realizarea celor doi algoirmi, unu care rezolvă o problemă de regresie și unu care soluționează clasificarea tumorii. Date existente în tabel provin de la imagini digitalizate a unei mase mamare, oferind detalii importante referitoare la celulele tumorale. În setul de date utilizat nu există valori lipsă. Pentru problema de regresie care are ca și scop final predicția cu o eroare căt mai scăzută a ariei tumorii, iesirea datelor a fost sub forma continua. Pentru partea de clasificare s-a folosit coloana numita diagonisis care continea litere, B insemnând benignă și M reprezentănd malignă. Aceste litere au fost convertite în numere, B fiind înlocuit de cifra 0 și M de cifra 1. Cu ajutorul acestei mapări am reușit sa creăm doua clase posibile, astfel construind algoritmul de clasificare.

Învățarea automată oferă posibilitatea a doua tipuri de învățare: supravegheată sau nesupravegheată. Diferențele dintre cele două sunt reprezentate de modul de învățare a algorimilor. La învățarea supravegheată, ieșirea este prezisă in funcție de caracteristicile intrării și seturile de date sunt divizate în set de antrenare și set de testare.[5] Pe baza setului de antrenare se realizează învățarea, iar pe cel de validare o să se facă compararea între ce se dorește să se obțină și ce a reușit algoritmul să prezică. Se folosește un set diferit pentru partea de validare, pentru că se dorește verificarea algoritmului pe un set diferit față de cel pe care și-a realizat procesul de antrenare. Algoritmi caracteristici învățării nesupravegheată au ca și particularitate faptul că au libertatea să învețe prin descoperire.[5]

Lucrarea aceasta se bazează pe o învățare supravegheată. Datele de intrare sunt străns legate de datele dorite la ieșire, astfel creându-se o relație care ajută algoritmul sa ofere rezultate cu o acuratețe crescută. În plus, s-a realizat și o împărțire a datelor de antrenare și de validare. Setul de date corespunzător antrenării reprezintă 80% din total, iar cel de testare reprezintă un procent de 20%. În general, pentru probleme de predicție si de clasificare este recomand folosirea învățării supravegheate.



Figură 2.1. Învățare supravegheată

# Analiză, proiectare, implementare

Aceasta parte a lucrării este flexibilă și depinde foarte mult de natura lucrării, poate fi organizată în mai multe capitole și conține contribuțiile personale ale autorului.

Includeți:

* + Detalii referitoare la analiză și proiectare:
    - descrierea metodelor pe care le-ați aplicat pentru rezolvarea problemei,
    - descrierea materialelor, procedurilor
    - calcule, tehnici, descrierea echipamentelor
    - metodologia de proiectare
    - informațiile necesare pentru ca cineva să poată reface lucrarea
  + Implementare :
    - Descrieți detaliile tehnice ale implementării aplicației: mediul de implementare, modul de prezentare, modul de utilizare al aplicației, etc.
  + Testare si validare :
    - Descrieți metodologia de testare a aplicației și rezultatele
    - Includeți experimentele pe care le-ați realizat, analiza rezultatelor pe care le-ați obținut.

# Concluzii

## Rezultate obținute

Evidențiați toate rezultatele pe care le-ați obținut și trageți concluzii din ele. Puteți prezenta o analiză critică a ceea ce ați realizat comparativ cu alte lucrări/studii anterioare.

Includeți o listă a contribuțiilor pe care le-ați avut în domeniul temei abordate.

## Direcții de dezvoltare

Descrieți direcțiile posibile de dezvoltare.

# Reguli de formatare

## Formatarea paginii

* + Dimensiunea paginii: A4
  + Margini: 2.5 cm (sus, jos, stânga, dreapta)
  + Antet și subsol: 1.27 cm de la marginea paginii
  + În antetul paginii (header): titlul capitolului, centrat, stil: Header\_style
  + În subsolul paginii: numărul paginii, centrat

## Titluri și stiluri

Titlurile capitolelor și subcapitolelor se marchează cu stilurile Heading 1 – 4, conform documentului model anexat în format Word. Descrierea stilurilor utilizate în document este prezentată în Tabelul 5.1.

Tabelul 5.1. Stiluri utilizate în acest document

| Nr. | Stil | Utilizat pentru | Format |
| --- | --- | --- | --- |
| 1 | Normal | Text normal | Font: (Default) Cambria, 12 pt, Justified, Line spacing: Multiple 1.1 li, Space After: 6 pt |
| 2 | Titlu | Titlul proiectului, prima pagină | Font: 24 pt, Small caps, Centered Line spacing: single, Space Before: 126pt, After: 0 pt, |
| 3 | Titlu2 | Titlul proiectului, pagina de prezentare | Font:14pt, Bold, Centered |
| 4 | Heading 1 | Titlurile capitolelor (nivel 1) | Font: 24 pt, Indent: Left: 0 cm Hanging: 0.76 cm, Space Before: 24pt, After: 12pt |
| 5 | Heading 2 | Titlurile subcapitolelor (nivel 2) | Font: 14 pt, Bold, Indent: Left: 0 cm  Hanging: 1.02 cm, Space Before: 18pt, After: 12pt |
| 6 | Heading 3 | Titlurile secțiunilor (nivel 3) | Font: Bold, Indent: Left: 0 cm Hanging: 1.27 cm, Space Before: 6 pt, After: 6pt |
| 7 | Heading 4 | Titlurile secțiunilor (nivel 4) | Font: Italic, Indent: Left: 0 cm Hanging: 1.52 cm, Space Before: 2 pt, After: 0 pt |
| 8 | Caption | Legenda figurilor și tabelelor | Font: Italic, Font color: Text 1, Line spacing: single, Space After: 10 pt, |
| 9 | Header\_style | Antetul paginii | Font: 10 pt, Italic, Centered, Border: Bottom: (Single solid line, Background 1, 0.5 pt Line width) |

## Figuri, tabele și ecuații

### Figuri

Figurile se inserează în text centrate, cu etichetă de numerotare și legendă (Caption) în partea de jos a figurii. Numărul figurii include și numărul capitolului, după exemplul prezentat în Figura 5.1.



Figura 5.1. Figură exemplu, stil: Caption

## Tabele

Tabelele se inserează în text centrate, cu etichetă și legendă (Caption) în partea de sus a tabelului, aliniată la stânga. Numărul tabelului include și numărul capitolului, după cum este prezentat, de exemplu, în Tabelul 5.1.

## Ecuații

Ecuațiile se inserează în text centrate, cu numerotare în partea dreaptă. Numărul ecuației include și numărul capitolului, conform exemplului din relația (5.1).

|  |  |
| --- | --- |
|  | (5.1) |

## Referințe bibliografice

Se recomandă ca citarea referințelor bibliografice să fie făcută în formatul IEEE.

În secțiunea Bibliografie sunt prezentate exemple pentru: o citare a unui capitol dintr-o carte [1], un articol publicat într-o revistă [2] și un articol publicat la o conferință [3].

Detalii cu privire la formatul citării diverselor tipuri de referințe pot fi găsite în [4] sau [5].

Referințele bibliografice se pot insera în text utilizând facilitățile Word de a adăuga surse și bibliografie unui document (References -> Citations & Bibliography). Dacă formatul IEEE pentru bibliografie nu este instalat implicit în Word, se poate descărca gratuit de la:

<https://bibword.codeplex.com/wikipage?title=Styles&referringTitle=Home>

Instrucțiunile de instalare pentru diferite versiuni de Word se pot obține de la aceeași adresă.

6.Bibliografie

[1] Yue, Wenbin, Zidong Wang, Hongwei Chen, Annette Payne, and Xiaohui Liu. 2018. "Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis" Designs 2, no. 2: 13.

[2] Patel A. Benign vs Malignant Tumors. JAMA Oncol. 2020;6(9):1488.

[3] Wolberg,William, Mangasarian,Olvi, Street,Nick, and Street,W.. (1995). Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository.

[4] ABRAHAM, Ajith. Artificial neural networks. *Handbook of measuring system design*, 2005

[5] Mahesh, Batta. (2019). Machine Learning Algorithms -A Review. 10.21275/ART20203995.