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 aproximativ 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](#_Toc168493693)

[1.1 Context general 2](#_Toc168493694)

[1.2 Obiective 3](#_Toc168493695)

[1.3 Specificații 4](#_Toc168493696)

[2 Studiu bibliografic 5](#_Toc168493697)

[3 Analiză, proiectare, implementare 14](#_Toc168493698)

[3.1 Mediu de dezvoltare 14](#_Toc168493699)

[3.2 Caracteristicile setului de date 14](#_Toc168493700)

[4 Concluzii 16](#_Toc168493701)

[4.1 Rezultate obținute 16](#_Toc168493702)

[4.2 Direcții de dezvoltare 16](#_Toc168493703)

[5 Reguli de formatare 17](#_Toc168493704)

[5.1 Formatarea paginii 17](#_Toc168493705)

[5.2 Titluri și stiluri 17](#_Toc168493706)

[5.3 Figuri, tabele și ecuații 18](#_Toc168493707)

[5.3.1 Figuri 18](#_Toc168493708)

[5.4 Tabele 18](#_Toc168493709)

[5.5 Ecuații 18](#_Toc168493710)

[6 Bibliografie 19](#_Toc168493711)

# Introducere

## Context general

Cancerul de sân a reprezentat dintotdeauna unul dintre principalele motive de deces în răndul femeilor, aproximativ 15% din numărul total de cazuri.[1] În anul 2020, statisticile au oferit un număr apropiat de 2.3 milioane de cazuri noi apărute în rândul persoanelor de sex feminin, dintre care au dus la decesele a 685.000 de femei. Se preconizează o creștere accelerată a numărului de cazuri, astfel în anul 2040, numărul personelor care o să dețină acest diagnostic o să depășească 3 milioane. Din această valoare, decesele o sa reprezinte o treime.[7] Se poate observa tendința de creștere a numărului de femei care se confruntă cu această neoplazie, rezultând o necesitate acută pentru modalități de cunoaștere și de înțelegere aprofundată a modului de expansiune a bolii.

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. Cu cât procedura medicală este administrată din timp, cu atât probabilitatea de supraviețuire a persoanelor se amplifică.

Prin această lucrare, s-au atins două puncte vitale care au rolul de a carateriza o tumoare: modul în care o să evolueze dimensiunea ariei tumorii, bazat pe anumite valori medicale, respectiv dacă aceasta are o caracteristică malignă sau benignă. Tumoarea reprezintă o acumularea excesivă de celule într-o zonă a organismului. Principalul motiv pentru care se formează fiind reprezentat de operațiunea celulelor de a se divide anormal de mult într-un timp scurt sau din simplu fapt că acestea nu mor într-un timp așteptat de către organism. Tumoarea benignă în mod normal nu cauzeaza împrejurări dificile pacienților, dar chiar dacă au un ritm lent de creștere, pot să ajungă în punctul în care să influențeze funcționarea normală și eficientă 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 în zona inițială, dar și să ajungă în tot organismul uman prin intermediul sângelui.[2] Aceste aspecte arată importanța cunoașterii tipului tumorii, din acest motiv, lucrarea prezintă procesul prin care s-a studiat și încercat găsirea unui algoritm cât mai potrivit pentru această sarcină. De sigur, cunoscând caracteristica tumorii maligne, creștere sporadică, se justifică și motivația medicilor de a cunoaște maniera de evoluție a mărimii afecțiunii. Această lucrare reflectă dorința de a ajuta și de a ușura modul de gestionare a tratamentului oncologic aplicat oamenilor. Cancerul reprezentând o problemă medicală cu o frecvență de expansiune înaltă în viața cotidiană, este foarte important orice detaliu suplimentar despre maniera de evoluție în organismul uman.

De-a lungul anilor, aria medicinei a cunoscut o dezvoltare cu un imens succes, astfel reușind să ofere tratamente adecvate și mult mai eficiente pentru pacienți. Totuși, învățarea automată poate să aducă un beneficiu considerabil în clasificarea tumorii, dar si pentru modul în care aria formațiunii o să avanseze pe parcursul timpului. [1] O astfel de predicție eficientă asupra particularității neoplasmului poate să îi ajute si pe specialiștii 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. S-au efectuat foarte multe studii pe toate sferele medicinei. Partea de cancer este una care reușește să ocupe un loc principal în majoritatea studilor și cercetărilor, deoarece pentru această afecțiune nu există un leac sau un tratament care să ofere o marjă crescută în cea ce privește vindecarea.

Inteligența artificială (AI) a devenit tot mai prezentă în viața de zi cu zi a oamenilor. Programatorii au început să o utilizeze pentru a putea oferi metode simplificate și cu o acuratețe crescută, în majoritatea problemelor existente. Fiind un domeniu în continuă dezvoltare și reprezentând o sferă de interes pentru cei mai mulți informaticeni, se încearcă răspândirea ei pe diverse domenii de activitate. Învățarea automată este un subdomeniu al inteligenței artificiale. Cu ajutorul acestei tehnologii se crează anumiți algoritmi care pot să învețe cum să ofere o predicție cât mai aproape de adevăr, pe baza unor seturi de date și relații între acestea.

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 de acumulare a informațiilor. Acesta este alcătuit dintr-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 nouă.[4]

Lucrarea de față oferă o soluție pentru medicii de pe secția de oncologie, să cunoască mai multe informații relevante la adresa tumorii. Acesta combină necesitatea medicală cu învățarea automată, astfel putând să se ajungă la un mod de lucru mai eficient. Cunoscând modul de evoluție a patologiei, cadrele din domeniul medical pot să aibă o idee despre cum o să evolueze stadiul bolii, astfel reușind să prescrie un tratament mult mai eficient.

În continuare sunt prezentate 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 neoplaziei. [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 pune la dispoziție și un algoritm de clasificare a tipul formațiunii cu o acuratețe crescută.

Predicția dimensiunii neoplasmului și clasificarea tipului acestuia, 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.

Toate aceste informații provenite de la mase mamare, au contribuit la realizarea rețelelor neuronale artificale, având 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, doctorii 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 rezultatele obținute în această documentație, 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 rata administrării unor 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 tratatementului, poate să ducă o la creștere de supraviețuire crescută. Calitatea îngrijirii medicale ar trebui să cunoască o îmbunătățire semnificativă, 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 și simplificată a problemelor bazate pe date.[5]

Ea utilizeaza elemente cu o capacitate de optimizare avansată față de procedurile existente în trecut. Are capacitatea să creeze relații și corelații între datele primite, astfel reușind să observe un tipar sau anumite caracteristici ale datelor de intrare. De-a lungul anilor s-au realizat diverse comparări între învățarea automată și diferite metode clasice de predicție, dar învățarea automată a reușit să ofere rezultate mult mai apropiate de adevăr.[1]

Învățarea automată oferă posibilitatea a trei tipuri de învățare: supravegheată, nesupravegheată sau semi-supravegheată. Diferențele dintre cele ele sunt reprezentate de modul de învățare a algorimilor. La învățarea supravegheată, ieșirea este prezisă în funcție de caracteristicile intrării, iar 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]

Învățarea semi-supravegheată este o combinație între cele două prezentate mai sus. Este potrivita pentru date care sunt etichetate, dar și ne-etichetate.[10]

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 să ofere răspunsuri 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.

În figura 1, se poate observa modul de funcționare a metodei de învățare supravegheată.

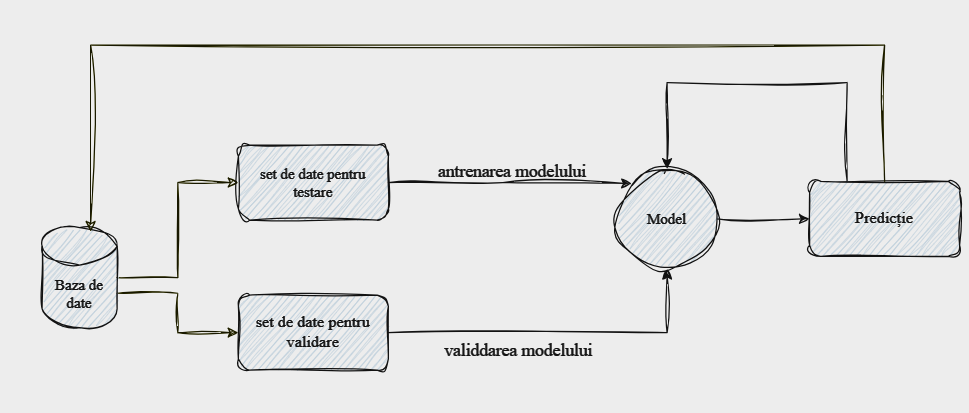


Figura 2.1. Învățare supravegheată

Datorată acestei împărțiri, învățarea automată a cunoscut apoi o noua fragmentare, astfel sunt anumite particularități ai algoritmilor care îi fac mai potriviți pentru rezolvarea unor probleme de regreie sau de clasificare. Clasificarea se bazează pe distribuirea valorilor de ieșire în anumite clase, practic se oferă o etichetă ieșirilor. [10]

Pentru a rezolva problema de clasificare a trebuit să se folosească metoda de codificare a etichetelor (label encoding). În setul de date utilizat, coloana în care se afla informația despre tumoare, dacă este malignă sau benignă, este prezentă sub formă de text. Pentru a putea lucra cu aceste date, a trebuit să fie convertite în numere.[8] Prin intermediul acestei tehnici, datele care descriau tumoare ca fiind malignă (M) sau benignă (B) au fost transformate în valorile de 0 sau 1. În acest mod, personele care citesc datele o să știe că în momentul în care văd valoarea de 0, tumoarea este benignă, iar pentru valoarea de 1, tumoarea este considerată periculoasă, malignă.

Pe de altă parte, regresia soluționează situații în care se dorește predicția unor valori continue [10]. În acest context, valorile care simbolizează aria tumorii sunt numere reale care nu pot să fie clasificate într-un anumit domeniu.

Prin utilizarea unei învățării supravegheate se pot utiliza mai mulți algoritimi avansați cum ar fi: rețele neuronale artificiale (ANN), decision tree (DT), support vector machine(SVM).

O rețea neuronală artificială este formată 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 cel de ieșire. Ultimul nivel 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 absoarbe un lucru nou.[1] Omul își crează diferite conexiuni între ce trebuie să memoreze și ce cunoaște deja. El ajunge să repete o informația pe care trebuie să o rețină până în momentul în care acesta este capabil să reproducă cu o acuratețe crescută informația de care are nevoie. Același lucru se petrece și într-o rețea neuronală artificială, cu ajutorul epocilor repetă conținutul. În acest proiect ambele scopuri au fost atinse prin realizarea a două rețele neuronale, una care rezolvă problema de regresie și una care se pretează pentru problema de clasificare.

În figura 2.2, este atașată structura de baza a unui ANN. Acesta prezintă un singur strat ascuns cu 4 neuroni, un strat de input de 3 neuroni și un strat care înfățișează ieșirea. Se pot oberva relațiile care se construiesc între neuronii de pe toate straturile rețelei. Fiecare linie trasată de la un neuron la altu, semnifică conexiunile pe care neuronii și le crează. Fiecare conexiune are ca și caracteristică o pondere care modifică semnalul trimis.[4]

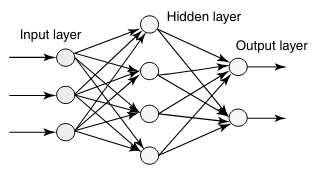


Figura 2.2 înfățișarea unui ANN

După momentul în care s-a construit rețeaua neuronală artificială, urmează pasul în care acesta este supusă proceselor de antrenare și de testare. Anterior rețelei, cum s-a pus în evidență mai sus, setul de date a trebuit să suporte și el un proces de prepocesare și de împărțire. Acesta poate să fie împărțit după cum consideră programatorul, dar și după felul în care modelul reușește să furnizeze o acuratețe suficient de mare. Regula ar fi că setul de date de antrenare ar trebui să fie mai mare sau egal cu cel de validare. Pentru acest proiect, modelul a funcționat cel mai bine și a reușit să ofere cele mai bune rezulatate la o împărtire de 80% antrenare și 20% testare.

Un obstacol foarte des întâlnit în problemele care utilizează învățarea automată, este evidențiat prin prezența conceptului de rețea neuronală artificială supra – antrenată sau sub–antrenată. În momentul în care o rețea este supra – antrenată, acesta pe parcursul procesului de antrenare memorează setul de antrenare, ajungând ca pe un alt set de date nou să aibă o eroare foarte mare. Practic aceasta nu își mai îmbunătățește în niciun fel abilitatea de a oferi o predicție bună și începe să se axeze prea mult pe detalile dintre date. Acest lucru este des întâlnit la seturile de date mici, pentru că modelul ajunge să țină minte toate caracteristicile setului mic de antrenare, iar în momentul în care se întâlnește cu un set nou (validare), nu o să poată să funcționeze corespunzător. [7]

Pe de altă parte, sub-antrenarea este și o dificultate care poate să apară în momentul antrenării. La supra-antrenare, modelul începea să memoreze datele pe care își efectua antrenarea, în loc să sustragă reguli. În cazul subantrenării, modelul nu reușește să capteze anumite reguli între date, prin care să ofere o predicție suficient de aproape de realitate. [7]

În figura 2.3, se creionează o comparație sugestivă între ambele situații distincte, supra- antrenarea și sub-antrenarea, dar include și forma corectă a unui model robust.



Figura 2.3 Stări ale rețelei

Un alt aspect notabil este numărul de straturi ascunse pe care o să le dețină rețeaua neuronală artificială. Fiecare strat ascuns adăugat o să conțină un număr de neuroni diferit. Cantitatea trebuie să fie descoperită în așa fel încât modelul să ajungă la o performanță apropiată de ideal. Dacă rețeaua are prea puține straturi ascunse, este posibil ca aceasta să nu fie capabilă să construiască suficiente legături cu ajutorul cărora să învețe să pună la dispoziție un rezultat conform realitații.[4] Importanța numărului de neuroni pe fiecare strat a fost exemplificată pe modelul realizat în soluționarea problemei, prin grafice reprezentative. În literatura se specifică faptul că ar în rețeaua neuronală artificială ar trebui să fie utilizate un minim de aproximativ 2 straturi ascunse.[11]

Modelul neuronal artificial este alcătuit din mai mulți hiperparametri care pot să fie modificați. Prin modificarea acestora se influențează maniera de funcționare a modelului. Schimbările oferă programatorului posibilitatea să observe cum este afectat modelul creat, reușind să găsească o combinație care să-l ajute să atingă o acuratețe ridicată. Spre exemplu, numărul de straturi ascunse simbolizează și el un hiperparametru al arhitecturii modelului.

În momentul antrenării modelului neuronal, se folosesc diferiți algoritmi de optimizare. Existe diferite tipuri de algoritmi de optimizare, câteva exemple sunt următoarele: Adam, RMSProp, SGD.

Algoritmul de optimizare Adam este unul dintre cei mai exploatați algoritmi de optimizare, în probleme care includ învățarea automată. El are mai multe avantaje, printre care și nevoia de memorie mică se numără. Oferă o recalculare a ratei de învățare pentru fiecare pondere a rețelei. [9]

RMSProp este tot un algoritm care este deseori în multe situații care implică modele neuronale artificiale. Acesta are o fromulă de calcul diferită pentru rata de învățare, față de cea prezentat în cazul algoritmului numit Adam. Metoda de calcul se bazează pe media pătratelor gradientelor. [9]

O alta metoda fezabilă de optimizare este descrisă de algoritmul SGD. El calculează funcția de pierdere pentru un singur eșantion într-un anumit moment specificat. Acesta nu ia în considerare tot setul de date de antrenare pentru a efectua calculul. Este un algoritm care este ales și preferat în multe probleme de lucru cu date.[9]

Fiecare algoritm de optimizare oferă o funcție unică de calculare a ratei de învățare. Persoana care construiește rețeaua neuronală trebuie să se informeze atent și să încerce diferite combinații între algoritmul de optimizare, arhitectura rețelei și hiperparametrii modelului, pentru a puteta ajunge la niște concluzii concludente. Pentru a putea observa care combinație conduce spre rezultate cu o eroare mult mai mică sunt necesare teste cât mai variante și observarea modificărilor pe grafic a valorilor.

Fiecărui algoritm de optimizare i se poate atribui o rată de învățare diferită. Aceasta poate juca un rol important în perfromanțele modelului. Există două cazuri, când rata de învățare este mult prea mică și momentul în care este prea mare. În cazul în care rata este mult prea mare, de fiecare dată minimul local o să fie ignorat, conducând spre rezultate nedorite. O să prezinte oscilații, dar și prezența unei grad lent care să îndrume spre o eroare mică. Pe de altă parte, o rată de învățare foarte mică poate să necesite un număr mare de epoci. Rețeaua ne fiind capabilă să acumeleze destule legături între date, astfel că ea nu reușește să ajungă la o eroare suficient de mică. Performanțele rețelei ar putea să ajungă să fie foarte lente, neputând să îndeplinească scopul final al modelului.[4]

Mai sus s-au specificat asemănările între modul de funcționare a creirului uman și procedeul prin care o rețea neuronală artificială acumulează informații. Pentru a oferi predicții cât mai corecte, algorimtii folosesc toate datele de intrare ca și caracteristici și încearcă să obțină niște reguli de mapare cât mai robuste. Există anumite date de intrare care sunt neliniare sau care oferă un grad de complexitate ridicat pentru maparea lor spre rezultatul dorit. Pentru soluționarea acestei dificultăți au fost create funcțiile de activare. Acestea sunt folosite pentru a limita amplitudinea valorilor de ieșire într-un număr întreg.[11]

Fiecare strat al rețelei neuronale artificiale are propria funcție de activare. Ea poate să difere în funcție de strat sau poate să rămână aceeași. Este vital să se înțeleagă faptul că se poate impune un prag de activare. Cu alte cuvinte, dacă datele de intare pentru funcția de activare respectă pragul impus, neuronul este considerat activ.[11]

Modul de lucru într-o rețea neuronală artificială este următorul: se iau intrările și se obține rezultatul sumei lor alături de greutățile lor. Pe această sumă se alocă o funcție de activare anterior aleasă, generând ieșirea corespunzătoare stratului implicit. Ieșirea este apoi distribuită mai departe spre următorul strat care așteaptă date.[11]

Câteva exemple de funcții de activare utilizate pentru diferite scopuri ar fi: Sigmoid, Tanh, ReLU, Leaky ReLU, ReLU parametrizat, Binary step function, funcția liniară.[11]

1. Funcția pasului binar

Este cea mai simplă funcție de activare atât ca și implemnetare în mediul de dezvoltare Python, dar și ca mod de funcționare. Dezavantajul pe care îl are este că nu oferă rezultate bune pentru problemele de clasificare pe mai multe clase. În figura 2.4 se află atât formula matematică a funcției, cât și reprezentarea grafică. [11]

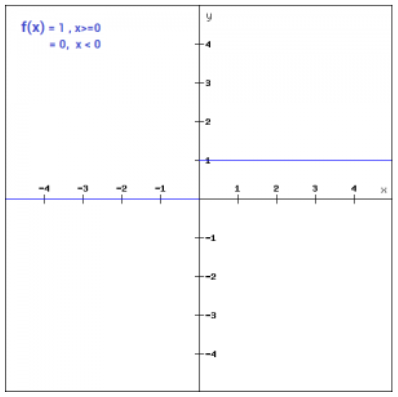


Figura 2.4 Pasul binar

1. Funcția sigmoid

Este una dintre cele mai utilizate pentru problemele de clasificare. Convertește rezultatul final în valori de 0 și 1. Semnele neuronilor vor fi consistente. În următoarea imagine este o reprezenatre explicită a funcției. [11]

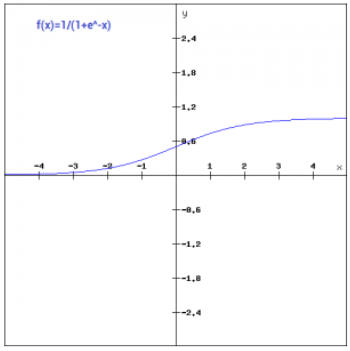


Figura 2.5 Funcția Sigmoid

1. Funcția liniară

Are o formă foarte ușoară. Aceasta este direct proporțională cu intrarea. Aceasta este ilustrată în figura 2.6. [11]

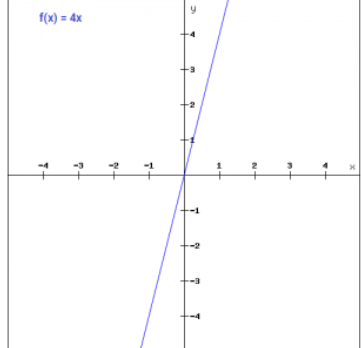


Figura 2.6 Funcția liniară

1. Funcția de activare ReLU

Se folosește în general în rețelele neuronale artificiale. Cel mai mare avatanj pe care îl are această funcție de activare este că neuronii nu sunt activați toți în același timp. Acest lucru contribuie foarte mult la creșterea eficienței în producerea de rezultate. O reprezenatre grafică a fost atașată mai jos pentru a oferi o înțelegere mai profundă. [11]

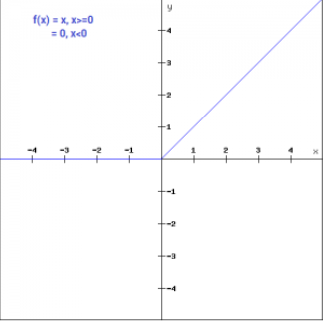


Figura 2.7 Funcția ReLU

1. Funcția Leaky ReLU

Este funcția relu construită pentru momentele în care x deține o valoare negativă, rezultând că o săa i se atribuie o valoare foarte mică. În momentul în care foloseam ReLU, dacă x dispunea de o valoare care are ca și caracteristică semnul minus, atunci funcția prelua valoarea 0. S-a atașat o imagine prin care se poate observa și asemănarea izbitoare cu funcția ReLU. [11]

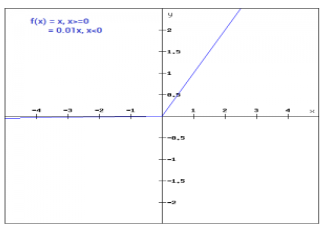


Figura 2.8 Funcția Leaky ReLU

1. Funcția Tahn

Diferența majoră între Tahn și Sigmoid, este subliniată prin faptul că această funcție de activare nu menține același semn pentru neuroni. Ampltudinea maximă și minimă a ieșiriilor este cuprinsă în intervalul 1 și -1. În figura 2.9 este exemplificată modelul grafic acestei funcții.

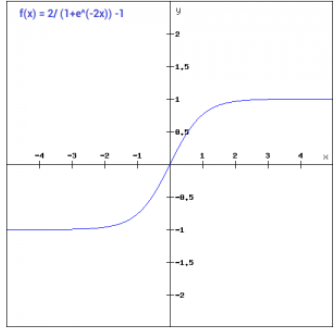


Figura 2.9 Funcția Tahn

Nu există anumite reguli stricte care să impună folosirea anumitor funcții de activare. Acest lucru depinde de datele de intrare, de ce se dorește să se obțină ca și rezultat, dar și de arthitectura rețelei. Ca și numărul de strauri sau numărul de epoci, neuroni de pe straturi, căutarea funcției de activare potrivită necesită o serie de teste asupra modelului. Totuși se cunoaște că Sigmoid oferă rezultate foarte bune pentru porblemele de clasificare, dar alături de Tahn, nu este recomandat să se folosească pe straturile din mijlocul rețelei. Ca și alternativă care oferă performanțe foarte bune, pentru straturile ascunse, este ReLu, iar Leaky ReLu în momentul în care există posibilitatea de neuroni morți. Acestea sunt numai niște constatări care au fost obținute în urma mai multor teste. [11] Toate aceste informații susțin faptul că funcția de activare are un rol semnificativ în rezultatele care vor fi înregistrate.

Pentru a putea constata dacă rețea oferă niște predicții corecte și apropiate de adevăr, pe lângă partea vizuală în care crează un grafic cu datele obținute prin predicție și datele reale păstrate pentru partea de tetstare, se utilizează diferite metrici de măsurare.

În mod firesc, există metrici de măsurare pentru probleme de clasificare, dar și speciale pentru soluționarea problemelor de regresie.

Pentru partea de clasificare, se utilizează trei metrice cunoscute: cea binară, multiclasă și multi etichetă. Cea binară oferă posibilitatea grupării rezultatului grupării în două opțiuni, 1 sau 0. Metoda mai multor clase acordă posibilitatea împărțirii ieșirii în mai mult de două clase, iar metoda mai multor etichete furnizează o sub clasificare a mai multor clase.

În lucrarea au fost folosite ca și metrici de măsurare a acurateței următoarele: eroarea medie pătratică (MSE) și coeficientul de determinare ().

Eroarea medie pătratică face parte din categoria numită media erorilor la pătrat. Această categorie este influențată mult mai ușor de valorile numerelor și se folosește în momentul în care se dorește o urmărire atentă a erorilor obținute prin predicție. Prin utilizarea acestei tip de erori se face o medie între ce se asteaptă să se obțină și ce s-a obținut de fapt. [13]

Coeficientul de determinare are un interval între 0 și 1. El arată cât de bine funcționează modelul. Cu cât este mai spre 1, cu atât modelul pune la dispoziție o predicție cu o acuratețe crescută. [13]

În tabelul 1 sunt trecute formulele de la cele două metrici. [13]

Tabel 2.1 Metrici de acuratețe

|  |  |  |  |
| --- | --- | --- | --- |
| Nr crt | Nume | Prescurtare | Formulă |
| 1 | Eroare medie pătratică | MSE |  |
| 2 | Coeficient de determinare |  |  |

Toate aceste informații au fost studiate în prealabil pentru a putea să se creeze două modele de rețele neuronale artificiale menite să ajute la soluționarea problemelor din domeniul medical, axate pe cancerul din zona mamară.

# Analiză, proiectare, implementare

Capitolul curent înglobează o descriere amănunțită a asupra bazei de date utilizate, mediului de dezvoltare, părți de cod relevante pentru a exemplifica modul de construire a rețelelor neuronale artificiale, dar și subcapitolul destinat testării.

Prin includerea acestei componente în documentație, se poate asigura o privire de ansamblu a modului de lucru abordat pentru soluționarea problemelor de regresie și de clasificare a tumorii mamare.

## Mediu de dezvoltare

Mediu de dezvoltare care a servit în rezolvare acestei lucrări este reprezentat de Pycharm Community Edition 2023. El pune la dispoziție posibilitatea navigării cu o dificulate scăzută prin proiect sau chiar sugestii la ce ar trebui programatorul să folosească în momentul scrierii codului.

Limbajul de programare folosit în cadrul acestui proiect, este Python cu versiunea de 3.9. s-a ales acest limbaj de programare, în ultima vreme reușind să câștige foarte mulți oameni dornici să-l utilizeze, pentru simplitatea cu care se poate scrie cod. Acesta oferă și o lizibilitate crescută și odată cu ea și înțelegerea codului devine ușoară.

Un alt aspect care a contribuit la alegerea limbajului, este dimensiunea largă a bibliotecilor prin care se realizează imprtul diferitelor funcții. Ele sunt reprezentate și de biblioteci specifice funcțiilor de activare a neuronilor, de creare a straturilor sau de calculare a unor metrici menite să ofere informații referitoare la perfomanțe.

## Caracteristicile setului de date

S-au folosit două seturi de date inițial: Breast Cancer Wistonsin (Diagnostic) [3] și Breast Cancer Wistonsin (Prognostic) [14]. În final s-a utilizat prima opțiune din motivul că cea de a doua punea la dispoziție un set de date mult prea mic. Ea îngloba 198 de instanțe. Au fost adăugate și câteva imagini pentru partea de regresie, în care este evidențiată importanța colecției de date și cât de mult infleunțează modul de învățare a rețelei. Cu un set de date mai mic, eroarea medie pătratică a putut să fie scăzută la o valoare de 0.03. Aceste lucruri sunt detaliate mai amănunțit în subcapitolul care înglobează părțile de testare.

Setul de valori asupra căruia s-au obținut cele mai bune rezultate în lucrarea prezentată, este găsit sub numele de Breast Cancer Wisconsin (Diagnostic), furnizând un set amplu de măsurători. Datele existente în tabel provin de la imagini digitalizate a unei mase mamare, oferind detalii importante referitoare la celulele care sunt implicate în construcția tumorii. Un detaliu fundamental este faptul că în nicio coloană nu oferă date lipsă. [3]

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 continuă și nu a presupun nicio modificare de format. Pe de alta parte, pentru problema de clasificare s-a folosit coloana numita diagonisis care continea litere, B insemnând benignă și M reprezentănd malignă. Pentru a putea lucra cu acestea a fost nevoie de utilizarea unei conversii, din text în cifre. Aceste litere au fost convertite în numere, B fiind înlocuit de valoarea 0, respectiv M fiind asociat cifrei 1. Cu ajutorul acestei mapări s-au creat două clase posibile, astfel construind algoritmul de clasificare.

Rețeaua neuronală artificială (ANN) necesită o divizare riguroasă a cantității. Ea impune o segmentarea a informațiilor pe care o să le utilizeze pentru procesul de antrenare și pentru procesul de testare. Vital este ca datele, atât de intrare cât și de ieșire, pe care se realizează antrenarea, să cuprindă cât mai multe informații. Modelul devine mai eficient care dispune de mai multe date pe care poate să creeze legături și neuronii să poată elabora relații puternice de înțelegere. Validarea este obilgatoriu să se realizeze pe date noi, nu pe cele pe care s-a realizat instruirea, astfel se observa comportarea modelului și se poate demonstra acuratețea reală a modelului într-un mediu necunoscut anterior. Prin această etapă se asigură că algorimii pot să fie folosiți în rezolvarea unor probleme din viața oamenilor. Metoda de divizare oferă mai multă credibilitatea asupra rezultatelor obținute și este capabilă să sublinieze diferitele erori pe care modelul ar putea să le întâmpine. S-a utilizat un procent de 80% destinată părții de antrenare și restul de 20% a fost alocate testării.

Preprocesarea datelor a constat și în utilizarea modului de normlizare sau standardizare. Modelul furnizat a oferit rezultate cu o îmbunătățire semnificativă cu ajutorul normlizării. Algorimii au reușit prin intermediul acestei metode să își mărească perfomanțele și acuratețea substanțial. S-au oferit niște valori mult mai fiabil, astfel obținăndu-se o eroare medie pătratică mult mai scăzută și un mod de lucru care oferă o capacitate de predicție ridicată. Acest aspect scoate în evidență importanța și modul de evoluție a tehnicilor folosite pentru obținerea rezultatelor finale, dar și necesitatea înțelegerii în profunzime a datelor cu care se lucrează. Fiecare cerință pentru care este construită rețeaua neuronală artificială, necesintând o anumită tehnică de utilizare a seturilor de date.

Pentru obținerea unei eficiențe suporite, primele aspecte de care ținut cont sunt: atenția la alegerea setului de date, segmentarea într-o porțiune pentru antrenare și una pentru validare și utilizarea unor tehnici de preprocesare utile. Aceștia sunt primii pașii pe care programatorul îi realizează în momentul în care se hotărăște să utilizeze învățărea automată pentru predicția anumitor aspecte.

## Rețelele neuronale artificiale

Proiectul de față are ca și scop principal oferirea unor soluții fezabile pentru medicii de pe secțile de oncologie din spitale.

S-au creat două rețele neuronale artificiale cu ajutorul limbajului de programare Python, astfel ele oferind o predicție suficinet de apropiată de realitate.

## Teste realizate pentru a ajunge la forma optimă

# 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) |

# 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. Machine learning algorithms-a review. International Journal of Science and Research (IJSR).[Internet], 2020, 9.1: 381-386

[6] Melina Arnold, Eileen Morgan, Harriet Rumgay, Allini Mafra, Deependra Singh, Mathieu Laversanne, Jerome Vignat, Julie R. Gralow, Fatima Cardoso, Sabine Siesling, Isabelle Soerjomataram, Current and future burden of breast cancer: Global statistics for 2020 and 2040

[7] JABBAR, H.; KHAN, Rafiqul Zaman. Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study). Computer Science, Communication and Instrumentation Devices, 2015, 70.10.3850: 978-981.

[8] LAVANYA, M.; PARAMESWARI, R. A multiple linear regressions model for crop prediction with adam optimizer and neural network mlraonn. International Journal of Advanced Computer Science and Applications, 2020, 11.4.

[9] REYAD, Mohamed; SARHAN, Amany M.; ARAFA, Mohammad. A modified Adam algorithm for deep neural network optimization. Neural Computing and Applications, 2023, 35.23: 17095-17112.

[10] Qifang Bi, Katherine E Goodman, Joshua Kaminsky, Justin Lessler, What is Machine Learning? A Primer for the Epidemiologist, *American Journal of Epidemiology*, Volume 188, Issue 12, December 2019, Pages 2222–2239

[11] SHARMA, Sagar; SHARMA, Simone; ATHAIYA, Anidhya. Activation functions in neural networks. Towards Data Sci, 2017, 6.12: 310-316.

[12] ERICKSON, Bradley J.; KITAMURA, Felipe. Magician’s corner: 9. Performance metrics for machine learning models. Radiology: Artificial Intelligence, 2021, 3.3: e200126.

[13] STEURER, Miriam; HILL, Robert J.; PFEIFER, Norbert. Metrics for evaluating the performance of machine learning based automated valuation models. *Journal of Property Research*, 2021, 38.2: 99-129.

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