

[Artificia](https://doi.org/10.1016/j.ailsci.2023.100061)l[IntelligenceintheLifeSciences3(2023)100061](https://doi.org/10.1016/j.ailsci.2023.100061)

|  |  |  |
| --- | --- | --- |
|  | Contentslistsavailableat[ScienceDirect](http://www.ScienceDirect.com) |  |
| ArtificialIntelligenceintheLifeSciences |
| journalhomepage:[www.elsevier.com/locate/ailsci](http://www.elsevier.com/locate/ailsci) |

ResearchArticle   
Machinelearningforlongitudinalmortalityriskpredictioninpatientswith malignantneoplasminSãoPaulo,Brazil   
GFSSilvaa,∗,LSDuarteb,MMShirassub,SVPeresc,MAdeMoraesb,AChiavegattoFilhoa a*SchoolofPublicHealth,UniversityofSãoPaulo,SãoPaulo,Brazil*   
b*SãoPauloStateHealthDepartment,DivisionofNon-communicablediseases,SãoPaulo,Brazil*   
c*SãoPauloStateHealthDepartment,OncocenterFoundationofSãoPaulo,SãoPaulo,Brazil*

|  |  |  |
| --- | --- | --- |
| article | info | abstract |
| *Keywords:*  Machinelearning  Artificialintelligence Predictivemodel  Cancer |  | Artificialintelligenceisbecominganimportantdiagnosticandprognostictoolinrecentyears,asmachinelearning algorithmshavebeenshowntoimproveclinicaldecision-making.Thesealgorithmswillhavesomeoftheirmost importantapplicationsindevelopingregionswithrestricteddatacollection,buttheirperformanceunderthis conditionisstillwidelyunknown.WeanalyzedlongitudinaldatafromSãoPaulo,Brazil,todevelopmachine learningalgorithmstopredicttheriskofdeathinpatientswithcancer.Wetesteddifferentalgorithmsusingnine separatemodelstructures.ConsideringtheareaundertheROCcurve(AUC-ROC),weobtainedvaluesof0.946 forthegeneralmodel,0.945forthemodelwiththefivemaincancers,0.899forbronchialandlungcancer,0.947 forbreastcancer,0.866forstomachcancer,0.872forcoloncancer,0.923forrectumcancer,0.955forprostate cancer,and0.917foruterinecervixcancer.Ourresultsindicatethepotentialofbuildingmodelsforpredicting mortalityriskincancerpatientsindevelopingregionsusingonlyroutinely-collecteddata. |

**1.Introduction**

Neoplasmsaredefinedbyabnormaltissuegrowthandcanbeclassi-fiedasbenignormalignant.Benign(noncancerous)neoplasmsarechar-acterizedbyslowandorganizedspread,thepresenceofwell-defined borders,andtheabsenceofaninvasivecharacteratboththetissueand organlevels.Malignant(cancerous)neoplasms,ontheotherhand,are characterizedbyoftenrapidanddisorganizedgrowth,withpoorlyde-finedbordersandpossibleinvasionofadjacenttissuesandorgans,im-plyingthepossibilityofmetastaticcancer[1,2].

AccordingtotheWorldHealthOrganization,about9.6millionpeo-plediedofcancerworldwidein2018,ofwhicharound70%werein middle-andlow-incomecountries[3].InBrazil,accordingtotheNa-tionalCancerInstitute(INCA)[4],about625,000newcaseswereex-pectedin2020,basedonestimatesfrombeforetheSARS-coV-2pan-demic,andweremainlydistributedamongcancersoftheprostate,fe-malebreast,colonandrectum,trachea,bronchusandlung,andstom-ach.Asforthetotalnumberofdeaths,accordingtotheBrazilianMor-talityInformationSystemdata(SIM),224,829deathswerecausedby malignantneoplasmsin2020,andthemostfrequentweretrachea, bronchus,andlung(28,516deaths),breast(18,032deaths),prostate (15,841deaths),stomach(13,850deaths),andcolon(12,422deaths) [5].

∗Corresponding[author.](mailto:gabriel8.silva@usp.br)

*E-mailaddress:*[gabriel8.silva@usp.br](mailto:gabriel8.silva@usp.br)(G.Silva).

ArtificialIntelligence(AI)hasbecomeanimportanttoolinthefield ofmedicine.Machinelearningalgorithmsarecapabletoidentifypat-ternsandtrendsfromdatathatmaynotbereadilyapparenttothe humaneye.Thisallowsmedicalprofessionalstomakemoreaccurate predictionsaboutpatientdiagnosisandprognosisandmakeinformed decisionsabouttheirtreatment.Machinelearninginhealthcarehasthe potentialtogreatlyimprovepatientoutcomesandmakethehealthcare systemmoreefficient.

GiventhegrowingscenarioofcancercasesinBrazilandaroundthe world[4,6],itisincreasinglyimportanttoimproveprognosticdecisions forcancerpatients[7].Theaimofthisworkistodevelopmachine learningalgorithmstopredicttheriskofdeathincancerpatientsin ordertoprovideinputsfortheirclinicalmanagement.

**2.Materialandmethods**

*2.1.Datasetdescription*

WeanalyzeddatacollectedfromtheHospitalCancerRegistry(RHC) oftheOncocenterFoundationofSãoPaulo(FOSP/SP)[8],apublicreg-istrythatmonitorspatientstreatedinthestateofSãoPaulosincethe year2000.RHChasinformationonthecancerdiagnosis,treatment, metastases,recurrences,ageandsexofthepatients,anddataonthe

<https://doi.org/10.1016/j.ailsci.2023.100061>  
[Received21December2022;Receivedinrevi](https://doi.org/10.1016/j.ailsci.2023.100061)sedform31January2023;Accepted1February2023   
Availableonline3February2023   
2667-3185/© 2023TheAuthors.PublishedbyElsevierB.V.ThisisanopenaccessarticleundertheCCBYlicense(<http://creativecommons.org/licenses/by/4.0/>)

*G.Silva,L.Duarte,M.Shirassuetal.*

healthfacilities/organizationswheretheconsultationswereperformed. Thedatasetincludesatotal99variablesand1085,380patientsfrom 2000throughSeptember2022.

Allvariablescollectedafterthecancerdiagnosisforeachpatient wereremoved.Thealgorithmsweretrainedwithtwelvevariables:sex, age,daysbetweenfirstphysicianvisitanddiagnosis,clinicalstageof cancer,categoryofmedicalservice,previousdiagnosis,typeofdiag-nosis,topographygroup,healthregionofresidence[9],morphology, healthinstitutionhabilitation,andhealthregionofdiagnosis.Thereare threelevelstothevariablecategorymedicalservice:1)privatecare, 2)publiccare,3)privatecare.Thevariablepreviousdiagnosispresents binaryinformation,1forpatientswhostartedlongitudinalfollow-up withapreviouscancerdiagnosisand0forpatientswithoutprevious diagnosis.Thevariabletypeofdiagnosispresentsfourcategories:1) clinicalexamination,2)non-microscopicauxiliaryresources,3)micro-scopicconfirmationand4)noinformation.Thevariablescancertopog-raphyandcancermorphologyarecategorizedwithICD-10andICD-O, respectively.Thevariablesrelatedtohealthregionhaveseventeendis-tinctvalues,referringtotheseventeenhealthregionsinthestateof SãoPaulo,Brazil.Thevariablehealthinstitutionhabilitationhasfifteen categories:1)HighComplexityOncologyCareUnit(UNACON),2)UNA-CONwithRadiotherapyService,3)UNACONwithHematologyService, 5)ExclusiveUNACONforPediatricOncology,6)HighComplexityOn-cologyCareCenter(CACON),7)CACONwithPediatricOncologySer-vice,8)GeneralHospitalwithOncologicalSurgery,9)UNACONwith RadiotherapyandHematologyServices,10)UNACONwithRadiother-apy,HematologyandPediatricOncologyServices,12)UNACONwith HematologyandPediatricOncologyServices,13)Volunteer,14)Inac-tive,15)ExclusiveUNACONforPediatricOncologywithRadiotherapy Service.Thefulldescriptionofthedatasetanditsvariablescanbefound inSupplementaryAppendixB(TableB1).

Onlypatientswithdiagnosesfrom2014to2017wereincluded, inordertoavoidlongerclinicaleffectsafterthediagnosis.Although thedatasetincludedasmallportionofpopulationfromotherBrazil-ianstates,welimitedthealgorithmdevelopmenttoresidentsofthe stateofSãoPaulo(93%oftotalpatients).Weincludedonlypatients withmalignantneoplasmsandexcludedcasesofnon-melanomaofthe skinastheyhadalowmortalityrate.Weanalyzedadultpatientsre-gardlessofsex.Thefinalsamplewascomposedofatotalof29,194 patients.

*2.2.Outcomedefinition*

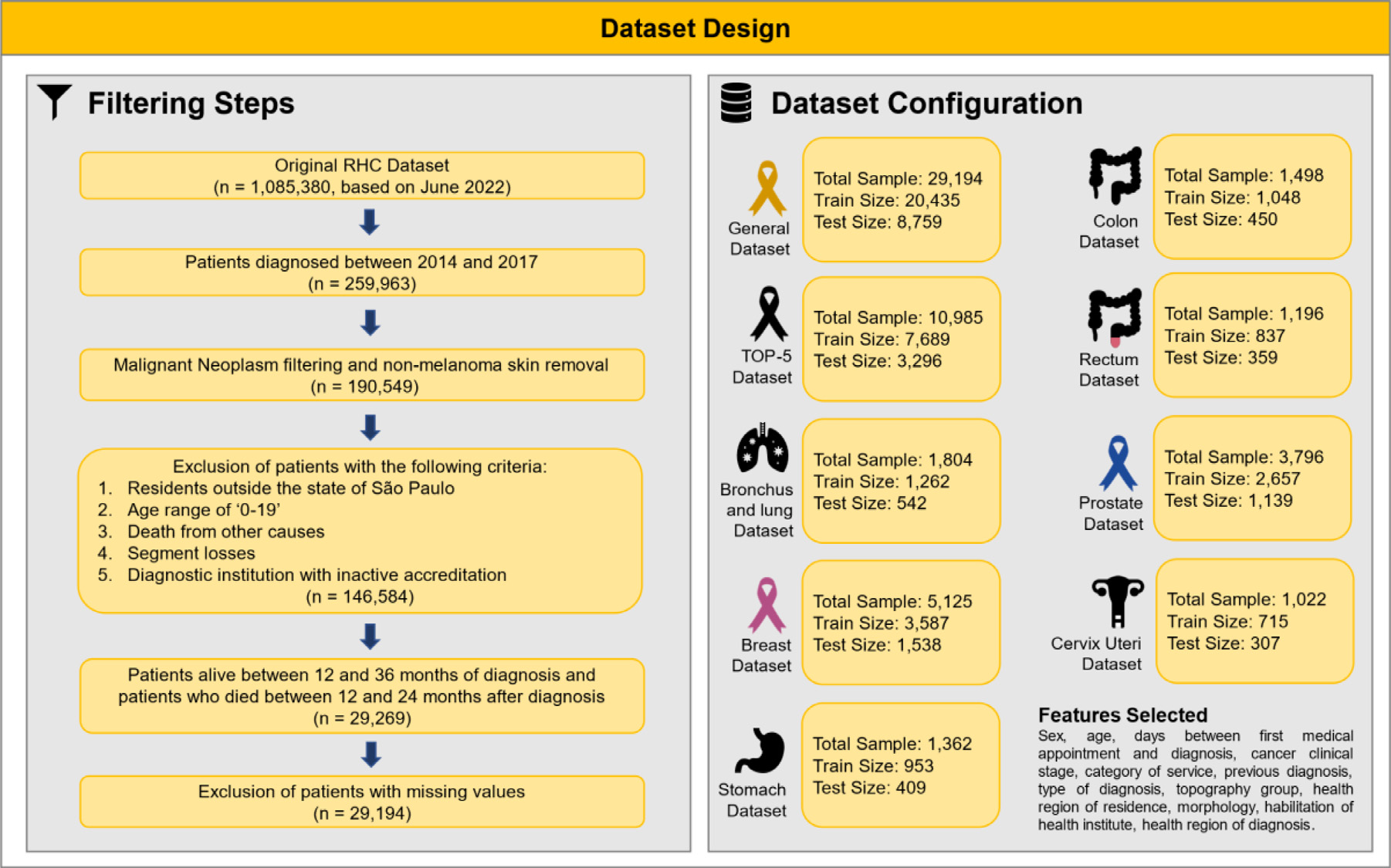
Theoriginaldatasetcontainsfourcategoriesregardingthelastavail-ableinformationaboutthepatient:1)alivewithoutcancer,2)alive withcancer,3)deathfromcancer,and4)deathwithoutfurtherinfor-mation.Ouroutcomeofinterestwaspatientswithaconfirmedcancer deathbetween12and24monthsafterthedateofdiagnosis.Forthe negativeoutcome,weincludedpatients1)alivewithoutcanceror2) alivewithcancerbetween12and36monthsafterthedateofdiag-nosis.Patientswereremovedifcategorizedas4)deathwithoutother information.

*2.3.Modeldesign*

Consideringthedistinctcancertypes,wetesteddifferentmodelsto assesswhetherchangingthestrategyincreasedmodelperformance.We firstdevelopedageneralmodelforallcancertypes.Wethendeveloped amodelforthetopfivecausesofcancermortality(bronchusandlung, breast,stomach,colon,andrectum).Wealsotrainedspecificmodels forthefivemostfrequentcausesandaddedtwoothermodelsbasedon itsgrowingimportanceforhealthvigilance:prostatecancerandcervix utericancer(inbothcasesthesexvariablewasnotusedasapredic-tor).Weevaluatedthemodelsindependently,withoutsharinganyin-formationduringalgorithmtraining.Asummaryofthemodeldesignis providedinSupplementaryAppendixA(FigureA1).

2

*G.Silva,L.Duarte,M.Shirassuetal.*  *ArtificialIntelligenceintheLifeSciences3(2023)100061*



**Fig.1.**Datasetfilteringandconfigurationsforpredictivemodels’development.

**Table1**

Descriptivesummaryoffull,trainandtestdatasets.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | FullDataset | Death | Non-death | Train | Test |
| **Sex**  Male  Female  **Age**  20–29  30–39  40–49  50–59  60–69  70+  **ClinicalStage**  I  II  III  IV  X  Y  **ServiceCategory** Public  Private  Particular  **utcome**  Death  Non-death | 14,313(49.0%) 14,881(51.0%) | 7027(55.2%)  5697(44.8%) | 7286(44.2%)  9184(55.8%) | 9972(48.8%)  10,463(51.2%) | 4341(49.6%) 4418(50.4%) |
| 929(3.2%)  2176(7.5%)  3862(13.2%)  7270(24.9%)  8093(27.7%)  6864(23.5%) | 263(2.1%)  636(5.0%)  1453(11.4%)  3234(25.4%)  3690(29.0%)  3448(27.1%) | 666(4.0%)  1540(9.4%)  2409(14.6%)  4036(24.5%)  4403(26.7%)  3416(20.7%) | 652(3.2%)  1534(7.5%)  2714(13.3%)  5077(24.8%)  5658(27.7%)  4800(23.5%) | 277(3.2%)  642(7.3%)  1148(13.1%) 2193(25.0%) 2435(27.8%) 2064(23.6%) |
| 6107(20.9%)  5917(20.3%)  5876(20.1%)  7602(26.0%)  712(2.4%)  2980(10.2%) | 570(4.5%)  1371(10.8%)  2843(22.3%)  6090(47.9%)  414(3.3%)  1436(11.3%) | 5537(33.6%)  4546(27.6%)  3033(18.4%)  1544(9.4%)  1512(9.2%)  298(1.8%) | 4292(21.0%)  4119(20.2%)  4088(20.0%)  5361(26.2%)  482(2.4%)  2093(10.2%) | 1815(20.7%) 1798(20.5%) 1788(20.4%) 2241(25.6%) 230(2.6%)  887(10.1%) |
| 21,224(72.7%) 7755(26.6%)  215(0.7%) | 11,865(93.2%) 803(6.3%)  56(0.4%) | 9359(56.8%)  6952(42.2%)  159(1.0%) | 14,840(72.6%) 5448(26.7%)  147(0.7%) | 6384(72.9%) 2307(26.3%) 68(0.8%) |
| 12,724(43.6%) 16,470(56.4%) | – – | – – | 8906(43.6%)  11,529(56.4%) | 3818(43.6%) 4941(56.4%) |

thenumberofpatientsinthetrainingandtestgroups.Therewereno-tableimbalancesaccordingtothedifferentmodels,with29,194patients inthegeneralmodeland1022inthecervixutericancermodel,which mayhaveaffectedthepredictiveperformanceofthealgorithms.

*3.3.Algorithmsperformance*

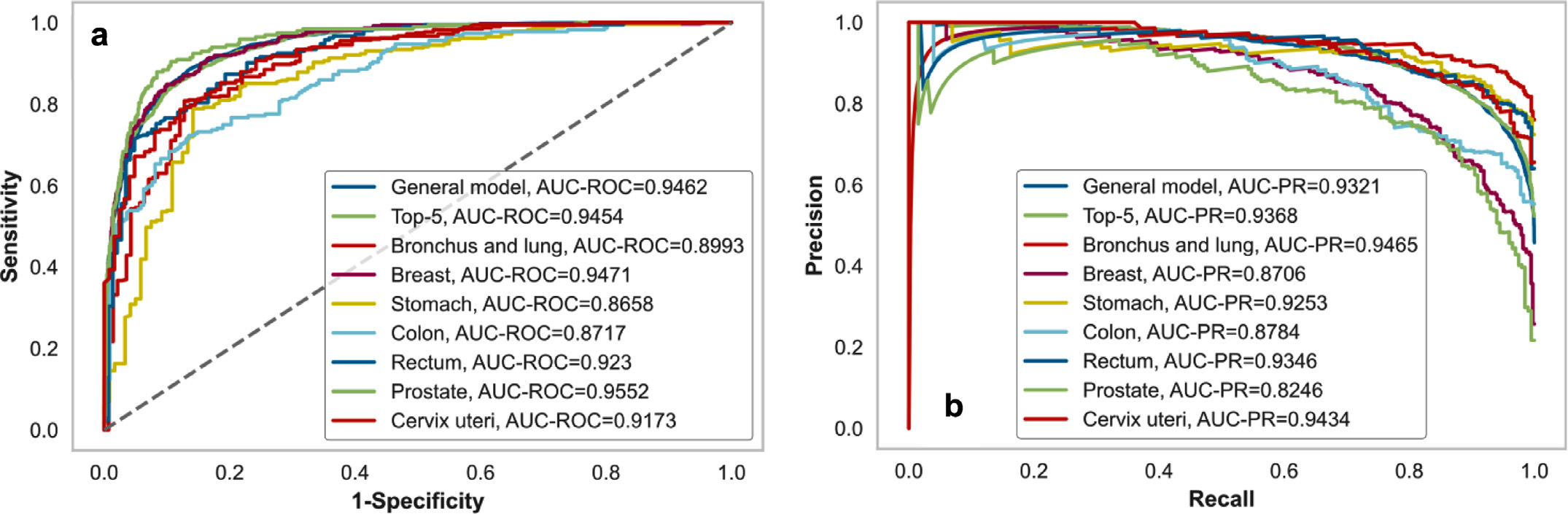
Thecatboostalgorithmpresentedthebestperformanceonallmodels exceptstomach,whereGradientBoostingperformedbetterintermsof AUC-ROC.Fig.2presentstheperformanceofthebestpredictionalgo-rithmsforeachofthemodelsconsideringtheAUC-ROCtestsetcriterion (a)andAUC-PR(b).Wefoundgooddiscriminationperformanceforthe

3

*G.Silva,L.Duarte,M.Shirassuetal.*  *ArtificialIntelligenceintheLifeSciences3(2023)100061*

**Table2**   
Descriptionoftheeightmodelsdevelopedforpredictionofcancermortality.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ID | Model | Variables | Totalcases | Non-death | Death | Mortality rate | TrainSize (70%) | TestSize (30%) |
| 1 | General | sex,age,medsvtodiag,cancerstage,servicecat,prevdiag, diagbase,topogroup,rras,morpho,habilit,rrasofserv  sex,age,medsvtodiag,cancerstage,servicecat,prevdiag, diagbase,topogroup,rras,morpho,habilit,rrasofserv  sex,age,medsvtodiag,cancerstage,servicecat,prevdiag, diagbase,rras,morpho,habilit,rrasofserv  sex,age,medsvtodiag,cancerstage,servicecat,prevdiag, diagbase,rras,morpho,habilit,rrasofserv  sex,age,medsvtodiag,cancerstage,servicecat,prevdiag, diagbase,rras,morpho,habilit,rrasofserv  sex,age,medsvtodiag,cancerstage,servicecat,prevdiag, diagbase,rras,morpho,habilit,rrasofserv  sex,age,medsvtodiag,cancerstage,servicecat,prevdiag, diagbase,rras,morpho,habilit,rrasofserv  age,medsvtodiag,cancerstage,servicecat,prevdiag,diagbase, rras,morpho,habilit,rrasofserv  age,medsvtodiag,cancerstage,servicecat,prevdiag,diagbase, rras,morpho,habilit,rrasofserv | 29,194 | 16,470 | 12,724 | 43.6% | 20,435 | 8.759 |
| 2 | Top-5causeof death  Bronchusand lung  Breast | 10,985 | 5937 | 5048 | 46.0% | 7689 | 3.296 |
| 3 | 1804 | 468 | 1336 | 74.1% | 1262 | 542 |
| 4 | 5125 | 3845 | 1280 | 25.0% | 3587 | 1.538 |
| 5 | Stomach | 1362 | 401 | 961 | 70.6% | 953 | 409 |
| 6 | Colon | 1498 | 740 | 758 | 50.6% | 1048 | 450 |
| 7 | Rectum | 1196 | 483 | 713 | 59.6% | 837 | 359 |
| 8 | Prostate | 3796 | 3232 | 664 | 17.5% | 2657 | 1.139 |
| 9 | Cervixuteri | 1022 | 416 | 609 | 59.6% | 715 | 307 |
| **medsvtodiag**:differenceindaysbetweenfirstmedicalappointmentdatesanddiagnosis,**cancerstage**:cancerclinicalstage,**servicecat**:categoryofservice,**prevdiag**: previousdiagnosis,**diagbase**:typeofdiagnosis,**topogroup**:cancertopography,**rras**:regionalnetofhealthcare(residence),**morpho**:cancermorphology,**habilit**: qualificationofthehealthestablishment,**rrasofserv**:regionalnetofhealthcare(service). | | | | | | | | |



**Fig.2.**PredictiveperformanceofbestalgorithmforeachmodelregardingAUC-ROC(a)andAUC-PR(b).

**Table3**   
Predictiveperformanceofbestalgorithmforeachmodel.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ID | Model | Best  Algorithm | Hypermeter Tunning | Feature  Selection | Resample | Accuracy | AUC-ROC | Recall | Specificity | Prec. | F1 | AUC-PR |
| 1 | General CatBoost Classifier Top-5cause CatBoost ofdeath Classifier Bronchus   CatBoost andlung Classifier Breast CatBoost Classifier Stomach Gradient Boosting Colon CatBoost Classifier Rectum CatBoost Classifier Prostate   CatBoost Classifier CervixuteriCatBoost Classifier | | None | None | None | 0.8743 | 0.9462 | 0.8549 | 0.8893 | 0.8565 | 0.8557 | 0.9321 |
| 2 | None | None | None | 0.8686 | 0.9454 | 0.8581 | 0.8776 | 0.8564 | 0.8572 | 0.9368 |
| 3 | None | None | SMOTE | 0.8561 | 0.8993 | 0.9152 | 0.6879 | 0.8929 | 0.9039 | 0.9465 |
| 4 | None | None | None | 0.8973 | 0.9471 | 0.7214 | 0.9558 | 0.8445 | 0.7781 | 0.8706 |
| 5 | RandomSearch None | None | None | 0.8093 | 0.8658 | 0.9343 | 0.5083 | 0.8207 | 0.8738 | 0.9253 |
| 6 | None | None | 0.7578 | 0.8717 | 0.7763 | 0.7387 | 0.7532 | 0.7646 | 0.8784 |
| 7 | Hyperopt | None | None | 0.8412 | 0.9230 | 0.9159 | 0.7310 | 0.8340 | 0.8731 | 0.9346 |
| 8 | None | None | None | 0.9210 | 0.9552 | 0.7487 | 0.9574 | 0.7884 | 0.7680 | 0.8246 |
| 9 | Hyperopt | None | None | 0.8306 | 0.9173 | 0.9235 | 0.6935 | 0.8164 | 0.8667 | 0.9434 |

4

*G.Silva,L.Duarte,M.Shirassuetal.*  *ArtificialIntelligenceintheLifeSciences3(2023)100061*

**Table4**

Predictiveperformanceofbestalgorithmforeachmodelbasedon20%individualswiththehighestriskofdeath.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ID | Model | TotalPatients | RealPositive | PositivePrediction | TruePositive | False  Positive | Precision | Recall |
| 1  2 | General  Top-5causeof death  Bronchusand lung  Breast  Stomach  Colon  Rectum  Prostate  Cervixuteri | 1749  659 | 1703  645 | 1749  659 | 1703  645 | 46  14 | 0.9737  0.9788 | 1.0000  1.0000 |
| 3 | 109 | 107 | 109 | 107 | 2 | 0.9817 | 1.0000 |
| 4  5  6  7  8  9 | 308  82  90  72  228  62 | 263  78  88  70  166  60 | 308  82  90  72  189  62 | 263  78  88  70  149  60 | 45  4  2  2  40  2 | 0.8539  0.9512  0.9778  0.9722  0.7884  0.9677 | 1.0000  1.0000  1.0000  1.0000  0.8976  1.0000 |

modelwastheonethatpresentedthebestAUC-ROC.WhenBORUTA wasused,therewasasignificantreductioninthenumberofpredictors (497to93)withoutalargelossinpredictiveperformance(AUC-ROCof 0.946fortherawmodelversus0.945forthemodelwithBORUTAand withouthyperparameteroptimization).Asimilarpatterntothegeneral modelwasobservedforthetop5causesofdeathmodel.Completere-sultsforalltrainingstrategiescanbefoundinSupplementaryAppendix B(TableB4).Thehyperparametersofeachmodelarealsoavailablein SupplementaryAppendixB(TablesB5,B6,B7,B8,B9,B10,B11,B12, B13).

Wealsoevaluatedtheperformanceofthealgorithmsinthetop20% (20%k-tops)ofpatientswiththehighestmortalityrisk(Table4).The generalmodelhad1749patientsinthegroup,ofwhom1703died,giv-ingthealgorithmaprecisionof97,37%andarecallof100%inthis high-riskgroup.Forthetop5causesofdeathmodel,659individuals wereinthe20%highestriskpatients,ofwhich645died,resultingina precisionof97.88%andarecallof100%.

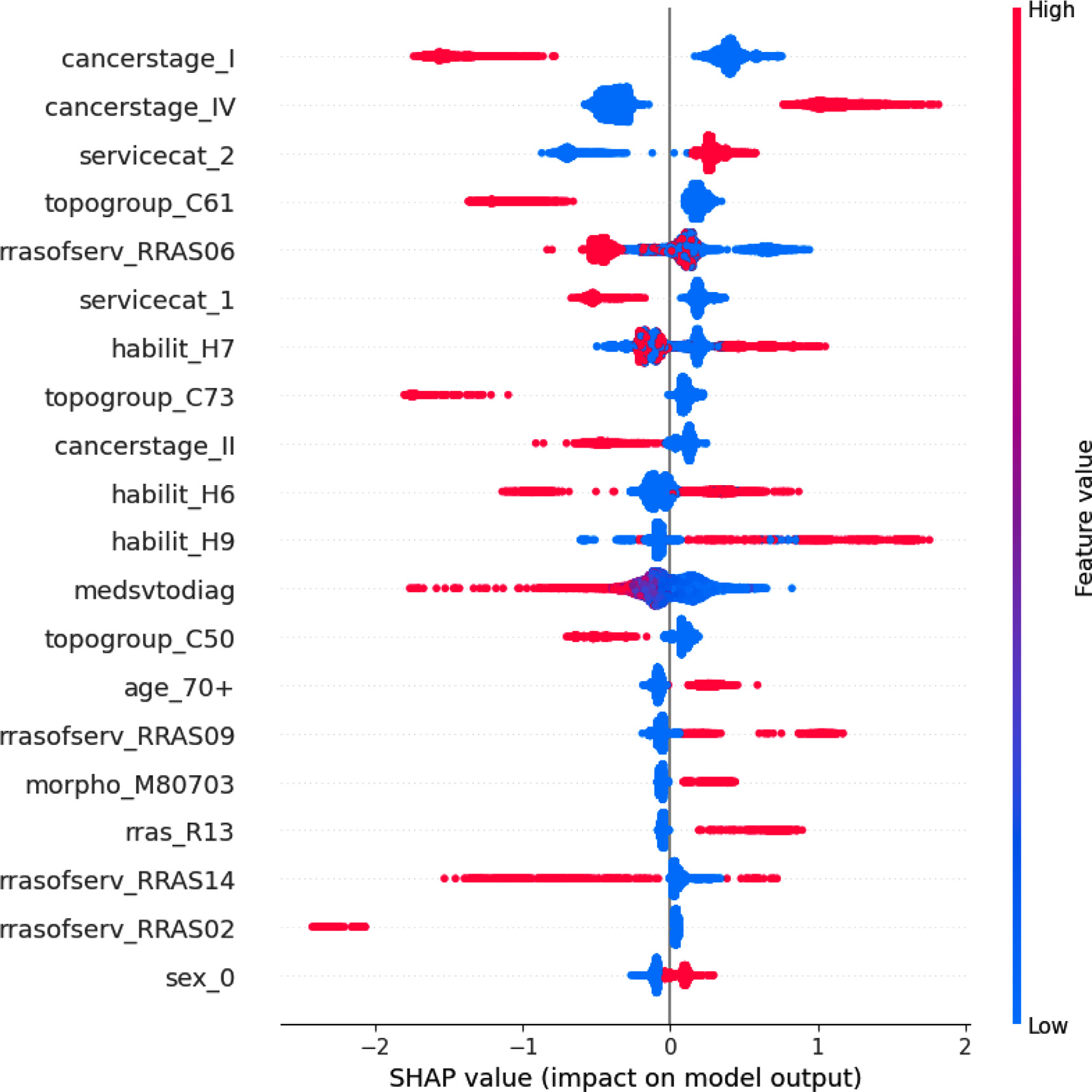
*3.4.Modelinterpretation*

Inordertointerpretthedecision-makingprocessofthealgorithms, wecalculatedtheShapleyvalues.Inthegeneralmodel(Fig.3),thecan-cerstageduringthefirstdiagnosiswasthemostimportantpredictor. StageIpatientsweremorelikelytobeclassifiednegatively(non-death), whereasstageIVpatientsweremoresignificantforthepositiveout-come(death).Thevariableonthecategoryofserviceprovidedwasalso importantfortheoutcome.Category2(publicservice)increasedmor-talityprediction,whereascategory1(privateservice)showedagreater propensityforpatientsurvival.Theothermainpredictivevariablesrefer tothetopographyofcancer,regionalnetofhealthcareservice(rrasof-serv)andregionalnetofhealthcareservice(rras).TheplotsofShapley valuesfortheothermodelscanbefoundinSupplementaryAppendixA (FiguresA2,A4,A6,A8,A10,A12,A14,A16).

Wealsorandomlyselectedthreepatients(highrisk,mediumrisk, andlowrisk)tohighlighttheindividualinterpretationofresults(Fig.4). Thefirst(a)wasatruepositive(riskof0.972)withtheexpectedShap-leyvaluewas2.92.Thevariablesthatcontributedfortheprediction ofpositiveoutcomewerepubliccareservice(servicecat\_2=1),cancer stagedifferentfromI(cancerstage\_*I*=0),cancertopographyICDC-34 (malignantneoplasmofbronchusandlung),regionalnetofservice01 (rrasofserv\_RRAS01=1)andthequalificationofthehealthcareinstitu-tion(code12,UNACONwithHematologyandPediatricOncologySer-vices).Asecondpatient(b)classifiedasafalsenegativewasselected. Thetotalriskscorewas0.4782,whichledthealgorithmtoclassifythe patientincorrectlyasaliveduringtheperiod.Weobservedthatthere wasbalanceintheaggregateofthecontributionofthepredictors,high-lightingtheimportanceofcancerstageIVtoincreaseShapleyvalue andthenon-publichealthservicetodecreaseit.Forpatientc,atrue negativeclassifiedaslowrisk,themostimportantcharacteristictoa lowexpectedShapleyvaluewerecancerstageIandnon-publichealth

5

*G.Silva,L.Duarte,M.Shirassuetal.*  *ArtificialIntelligenceintheLifeSciences3(2023)100061*



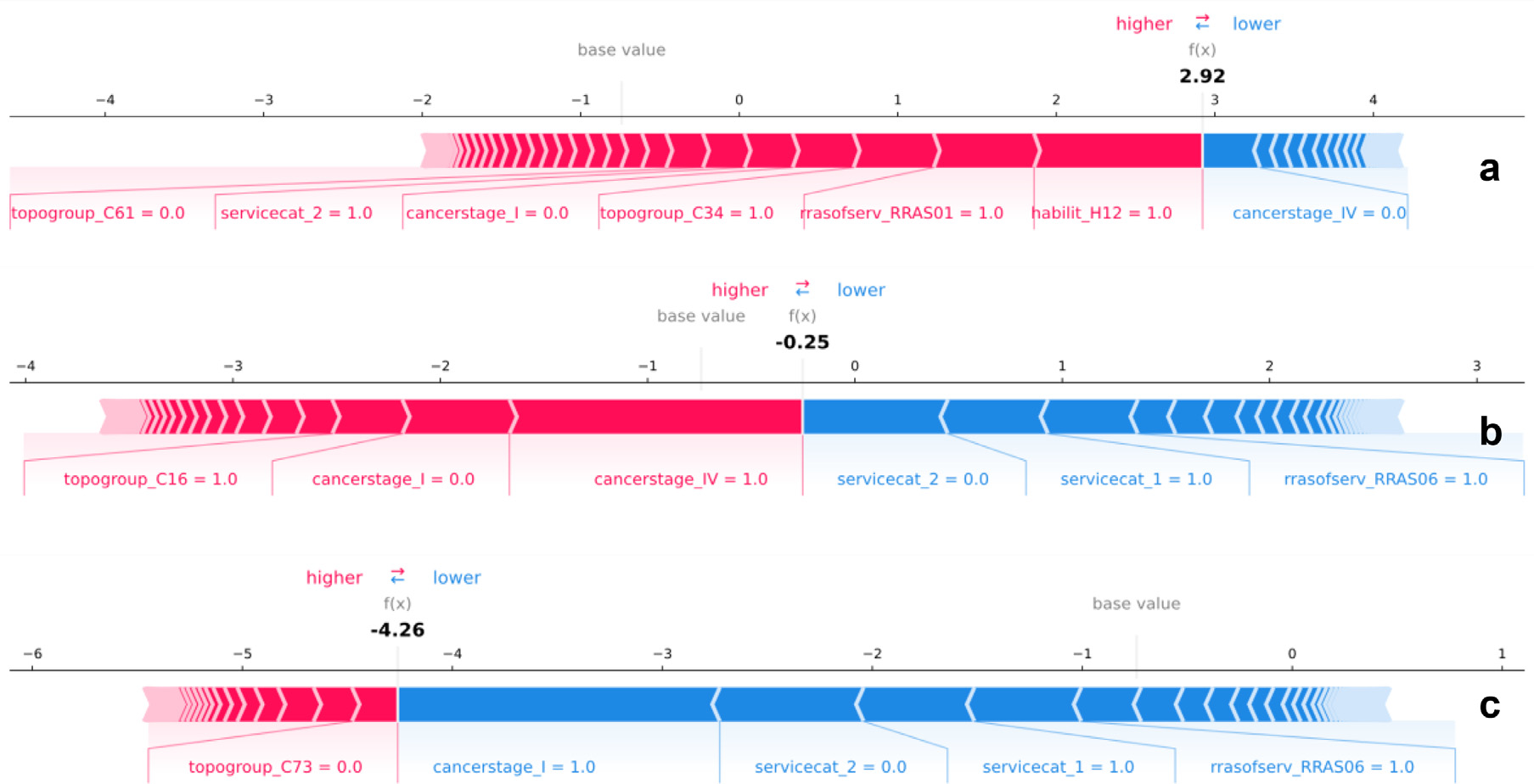
**Fig.3.**Toptwentypredictorsofriskofdeathfromcancer12to24monthsafterdiagnosis.GeneralModel,withCatboostClassifier.**cancerstage\_I**:cancerstageI,**can-cerstage\_IV**:cancerstageIV,**servicecat\_2**:publiccareservice,**topogroup\_C61**:cancertopographyICDC-61(malignantneoplasmofprostate),**rrasofserv\_RRAS06**: regionalnetofhealthcare(service)06,**servicecat\_1**:privatecareservice,**habilit\_H7**:qualificationH7CACONwithPediatricOncologyService,**topogroup\_C73**:can-certopographyICDC-73(malignantneoplasmofthyroidgland),**cancerstage\_II**:cancerstageII,**habilit\_H6**:qualificationH6CACON,**habilit\_H9**:qualificationH9 9-UNACONwithRadiotherapyandHematologyServices,**medsvtodiag**:differenceindaysbetweenfirstmedicalappointmentdatesanddiagnosis,**topogroup\_C50**: cancertopographyICDC-50(malignantneoplasmofbreast),**age\_70+**:agegroupof70yearsormore),**rrasofserv\_RRAS09**:regionalnetofhealthcare(service)09, **morpho\_M80703**:cancermorphology80,703(squamouscellcarcinoma,NOS),**rras\_R13**:regionalnetofhealthcare(residence)13,**rrasofserv\_RRAS14**:regional netofhealthcare(service)14,**rrasofserv\_RRAS02**:regionalnetofhealthcare(service)02.

**Table5**   
Comparisonofthepredictiveperformancebetweenthespecificalgorithmsforeachtypeofcancerandthegeneralalgorithm.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| CancerType | Model | Testsize | RealPositive | TruePositive | FalsePositive | TrueNegative | FalseNegative | Precision | Recall | AUC-ROC |
| Bronchusand Lung  Breast | General Specific General Specific General Specific General Colon  General Colon  General Specific General Specific | 486  542  1192  1538  383  409  468  450  319  359  1192  1139  326  307 | 367  401  206  384  277  289  241  228  197  214  206  199  215  183 | 340  367  146  277  258  270  200  177  177  196  146  149  177  170 | 35  44  29  51  26  59  36  58  26  39  29  40  21  42 | 84  97  957  1103  80  61  191  164  96  106  957  900  90  82 | 27  34  60  107  19  19  41  51  20  18  60  50  38  13 | 0.9067  0.8929  0.8343  0.8445  0.9085  0.8207  0.8475  0.7532  0.8719  0.8340  0.8343  0.7884  0.8939  0.8164 | 0.9264 0.9152 0.7087 0.7214 0.9314 0.9343 0.8299 0.7763 0.8985 0.9159 0.7087 0.7487 0.8233 0.9235 | 0.9265  0.8993  0.9460  0.9471  0.9255  0.8658  0.9241  0.8717  0.9163  0.9230  0.9460  0.9552  0.8850  0.9173 |
| Stomach |
| Colon |
| Rectum |
| Prostate |
| CervixUteri |

6

*G.Silva,L.Duarte,M.Shirassuetal.*  *ArtificialIntelligenceintheLifeSciences3(2023)100061*



**Fig.4.**Mainpredictorsofriskofdeathfromcancerbetween12and24monthsafterdiagnosisforthreerandomlyselectedindividuals:a)highriskofdeath(true positivewith0.972score),b)mediumriskofdeath(falsenegativewith0.478score)andc)lowriskofdeath(truenegativewith0.017score),generalmodelwith CatboostClassifier.**Patienta)topogroup\_C61:**cancertopographyICDC-61(malignantneoplasmofprostate),**servicecat\_2:**publiccareservice,**cancer\_stage1:** cancerstageI,**topogroup\_C34:**cancertopographyICDC-34(malignantneoplasmofbronchusandlung),**rrasofserv\_RRAS01:**regionalnetofhealthcare(service) 01,**habilit\_H12:**qualificationH12UNACONwithHematologyandPediatricOncologyServices,**cancerstage\_IV:**cancerstageI.**Patientb)servicecat\_1:**privatecare service,**morpho\_80,703:**cancermorphology80,703(squamouscellcarcinoma,NOS),**cancerstage\_I:**cancerstageI,**cancerstage\_IV:**cancerstageIV,**servicecat\_2:** publicprivatecareservice,**rrasofserv\_RRAS06:**regionalnetofhealthcare(service)06.**Patientc)topogroup\_C73:**cancertopographyICDC-73(malignantneoplasm ofthyroidgland),**cancerstage\_I:**cancerstageI,**servicecat\_2:**publiccareservice,**servicecat\_1:**privatecareservice,**rrasofserv\_RRAS06:**regionalnetofhealthcare (service)06.Zerovalueareinterpretedastheabsenceofthecharacteristicandoneasthepresence.

**4.Conclusion**  **Dataavailability**

Inconclusion,theninefinalmodelsdevelopedforpredictingriskof deathincancerpatientspresentedhighpredictiveperformance.Theal-gorithmscanbeanimportanttooltohelpprioritizetreatmentdecisions andpatientallocationincancertreatments,especiallyinlow-incomere-gions.Futureworkshouldexploretheproposedmethodologicalstruc-tureandevaluateitspredictiveperformanceinnewsettingswithdiffer-entroutinelycollecteddata.

**Ethicalstatement**

ThisworkwasevaluatedandapprovedbytheResearchEthicsCom-mitteeoftheFacultyofPublicHealthoftheUniversityofSãoPaulo (CAAE:65,375,722.9.0000.5421)

**Dataandcodeavailability**

Themainresultsofthisresearchwerepublishedinthisarticleand [insupplementaryappendixAandB.TheRHC/FOSPdataarepublicly availableinhttps://www.fosp.saude.sp.gov.br/fosp/diretoria-adjunta-de-informacao-e-epidemiologia/rhc-registro-hospitalar-de-cancer/.The codedevelopedforpredictivemodelingcanbeobtaineduponrequest.](https://www.fosp.saude.sp.gov.br/fosp/diretoria-adjunta-de-informacao-e-epidemiologia/rhc-registro-hospitalar-de-cancer/)

**Funding**

ThisworkwassupportedbytheSãoPauloStateHealthDepartment, SpecialHealthFundforMassImmunizationandDiseaseControl(FES-IMA/SES/SP).

**DeclarationofCompetingInterest**

Theauthorsdeclarethattheyhavenoknowncompetingfinancial interestsorpersonalrelationshipsthatcouldhaveappearedtoinfluence theworkreportedinthispaper.

7

*G.Silva,L.Duarte,M.Shirassuetal.*

[11]ChenT,GuestrinC.XGBoost:AScalableTreeBoostingSystem.In:Proceedingsofthe 22ndACMSIGKDDInternationalConferenceonKnow[ledgeDiscoveryandDataMi](https://doi.org/10.1145/2939672.2939785)n- [ing.NewYork,NY,USA:ACM;2016.p.785–94.doi](http://refhub.elsevier.com/S2667-3185(23)00005-3/sbref0012)[:10.1145/2939672.2939785.](https://doi.org/10.1145/2939672.2939785) [[12]KeG,MengQ,FinleyT,WangT,ChenW,MaW,LiuT-](http://refhub.elsevier.com/S2667-3185(23)00005-3/sbref0012)[Y.Lightgbm:Ahighlyeffic](https://doi.org/10.1145/2939672.2939785)[ient](http://refhub.elsevier.com/S2667-3185(23)00005-3/sbref0012)  [gradientboostingdecisiontree.AdvNeurInformProcessSyst2017;30:3146–54. [13]PedregosaF,VaroquauxG,GramfortA,MichelV,ThirionB,GriselO,DuchesnayE.](http://refhub.elsevier.com/S2667-3185(23)00005-3/sbref0013)  [Scikit-learn:MachinelearninginPython.JMachLearnRes2011;12:2825–30.](http://refhub.elsevier.com/S2667-3185(23)00005-3/sbref0013)

[[14]BergstraJ,YaminsD,CoxDD.MakingaScienceofModelSearch:Hyperparameter](http://refhub.elsevier.com/S2667-3185(23)00005-3/sbref0013) [OptimizationinHundredsofDimensionsforVisionArchitectures.In:Toappearin Proc.ofthe30thInternationalConferenceonMachineLearning(ICML2013);2013. [15]KursaMB,RudnickiW](http://refhub.elsevier.com/S2667-3185(23)00005-3/sbref0014)[R.FeatureSelectionw](https://doi.org/10.18637/jss.v036.i11)[iththeBorutaPackage.JStatSoftw](http://refhub.elsevier.com/S2667-3185(23)00005-3/sbref0014) [2010;36(11):1–13.doi](http://refhub.elsevier.com/S2667-3185(23)00005-3/sbref0016)[:10.18637/jss.v036.i11.](https://doi.org/10.18637/jss.v036.i11)

[[16]LundbergS,LeeS-I.](http://refhub.elsevier.com/S2667-3185(23)00005-3/sbref0016)A[unifie](https://doi.org/10.18637/jss.v036.i11)d[approachtoi](https://doi.org/10.18637/jss.v036.i11)[nterpretingmodelpredictions.NIPS;](http://refhub.elsevier.com/S2667-3185(23)00005-3/sbref0016)  [2017.](http://refhub.elsevier.com/S2667-3185(23)00005-3/sbref0016)

[[17]LundbergSM,NairB,VavilalaMS,etal.Explainablemachine-learningpredictions](http://refhub.elsevier.com/S2667-3185(23)00005-3/sbref0016) for[thepreventionofhypoxaemi](https://doi.org/10.1038/s41551-018-0304-0)aduringsurgery.NatBiomedEng2018;2:749–60. doi[:10.1038/s41551-018-0304-0.](https://doi.org/10.1038/s41551-018-0304-0)

8