UNIVERSITAT POLITÈCNICA DE VALÈNCIA

MIARFID
Máster en Inteligencia Artificial, Reconocimiento de Formas e Imagen Digital

Universitat Politècnica de València
Departament de Sistemes Informàtics i Computació

Aproximacions per donar suport al diagnòstic en base a imatge mèdica i altres dades clíniques mijançant tècniques d'aprenentatge profund

TREBALL FI DE MÀSTER

Màster Universitari en Intel·ligència Artificial, Reconeixement de Formes i Imatge Digital

Autora: Nadal Almela, Silvia

Tutors: Gómez Adrián, Jon Ander
Paredes Palacios, Roberto

Curs 2019-2020

Resumen

El objetivo principal de este proyecto es el desarrollo de herramientas software basadas en técnicas de aprendizaje automático, principalmente, redes neuronales, para asistir en la detección y diagnóstico de ciertas enfermedades. Se utilizarán conjuntos de imágenes médicas y, posiblemente, de datos complementarios a estas, para entrenar y ajustar los modelos diseñados. Los dos fines principales de los modelos serán clasificación, según si se presentan signos de la enfermedad o no, y segmentación, para resaltar aquellas zonas más interesantes de cara a determinar si el paciente sufre la enfermedad. Las imágenes se obtendrán de conjuntos de datos públicos, como el proporcionado por el Banco de Imagen Médica de la Comunidad Valenciana, formado por radiografías de pecho.

Palabras clave: Diagnóstico asistido, procesamiento de imagen médica, deep learning, redes neuronales

Resum

L'objectiu d'aquest projecte és el desenvolupament d'eines basades en tècniques d'aprenentatge automàtic, principalment, xarxes neuronals, per a assistir en la detecció i diagnòstic de certes malalties. Es farà ús de conjunts d'imatges mèdiques i, possiblement, de dades complementàries a aquestes, per a entrenar i ajustar els models dissenyats. Les dos principals finalitats dels models seran classificació, segons si es presenten signes de la malaltia o no, i segmentació, per a ressaltar aquelles zones més interessants de cara a determinar si el pacient pateix la malaltia. Les imatges s'obtindran de conjunts de dades públics, com aquell proporcionat pel Banc d'Imatge Mèdica de la Comunitat Valenciana, format per radiografies de pit.

Paraules clau: Diagnòstic assistit, processament d'imatge mèdica, deep learning, xarxes neuronals

Abstract

The main goal of this project is the development of software tools based on machine learning techniques, mainly neural networks, to assist in the detection and diagnosis of certain diseases. Sets of medical images and possibly complementary data will be used to train and adjust the designed models. The two main purposes of the models will be classification, depending on whether the images show signs of the disease, and segmentation, to highlight those areas that are more interesting to determine if the patient suffers from the disease. The images will be obtained from public data sets, such as that provided by the Medical Image Bank of the Valencian Community, consisting of chest X-ray images.

Key words: Support to diagnosis, medical image processing, deep learning, deep neural networks

Índex

Índex de figures

Índex de taules

CAPÍTOL 1
Introducció

Molts radiòlegs actualment han de llegir desenes d'estudis de rajos X al dia, la qual cosa suposa una càrrega de treball considerable. Per aquest motiu, es dissenyen ferramentes automàtiques que s'entrenen per a predir el risc d'anomalies específiques donada una imatge de rajos X particular. Aquestes ferramentes tenen el potencial per a fer front a la càrrega de treball dedicat a lectures de rajos X del radiòleg, ja siga augmentant la confiança del radiòleg com prioritzant per a la seua lectura aquelles imatges que corresponen a casos crítics. Aquestos són exemples de com es poden dissenyar sistemes de suport al diagnòstic –*Decision support systems*, DSS– com a ferramentes per a assistir en la interpretació clínica d'imatges de rajos X per a suplir unes necessitats no satisfetes.

Els models de Deep Learning per a classificació de patologies a partir de rajos X desenvolupats fins a l'actualitat no són generalitzables entre institucions i encara no estan preparats per a ser adoptats en escenaris reals

Entre els problemes que apareixen en voler aplicar Deep Learning en escenaris sanitaris, ens trobem amb la falta de dades de qualitat preparades per al seu ús. Al contrari que en conjunts d'imatges de domini general, com és ImageNet, les imatges de rajos X per a detectar patologies no poden ser anotades per grans grups de gent sense necessitat d'estar formades. Aquestes imatges han de ser anotades per experts en medicina. A açò se li suma el problema d'extraure informació de les anotacions dels metges mitjançant tècniques de processament de llenguatge natural –*Natural Language Processing*, NLP–. Aquestes anotacions solen reduir-se a un xicotet nombre de variables, o entitats, extretes de forma automàtica a partir del text. Aquesta reducció del *ground-truth* de la imatge sol contindre omissions i inconsistències i, a més, no és validada per cap professional mèdic.

1.1 Motivació

Aquest treball naix de la motivació per trobar solucions a la saturació que pateix el nostre sistema sanitari. La tecnologia actual permet crear propostes que aprofiten dades ja existents per a ajudar als professionals a prendre decisions. A més, ens trobem actualment vivint una pandèmia, la qual du més al límit la saturació sanitària ja mencionada. Per aquest motiu, s'ha decidit centrar aquest treball en

la detecció de COVID-19 mitjançant l'anàlisi automàtic d'imatges de pulmons. Gràcies al Banc d'Imatge Mèdica de la Comunitat Valenciana –BIMCV–, podem fer ús d'imatges provinents del nostre territori.

1.2 Objectius

Els objectius plantejats en aquest projecte són els següents:

- Crear models de xarxes neuronals aprofitant les imatges del BIMCV que siguen capaços d'etiquetar les mostres entre les tres classes possibles per a aquest dataset: control, pneumonia comuna i pneumonia provocada pel coronavirus.
- Comparar els resultats que obtenim utilitzant aquestes dades amb aquells resultats publicats en classificació de radiografies segons la presència o no de pneumonia i Covid-19.
- Aplicar mètodes d'*explicabilitat* sobre les prediccions proporcionades pels models amb la finalitat d'oferir un aclariment respecte aquestes prediccions.

1.3 Estructura del document

La memòria d'aquest projecte s'ha estructurat en sis parts principals. Primerament, es farà un resum de l'estat de l'art actual relacionat amb l'aplicació de tècniques d'aprenentatge automàtic que fan ús d'imatges mèdiques i, més concretament, de les investigacions al voltant de la COVID-19. A continuació, s'explicaran el problema i la solució proposada, detallant els mètodes i ferramentes emprats. Seguidament, es descriurà el conjunt de dades utilitzats. Després, s'analitzaran els resultats obtinguts. Finalment, es comentaran les conclusions a què s'ha arribat durant l'experimentació.

CAPÍTOL 2
Estat de l'Art

2.1 Aprenentatge automàtic en imatge mèdica

Les imatges mèdiques tenen diversos trets pels quals les solucions d'aprenentatge profund resulten idònies per a treballar amb elles. Entre aquestes qualitats es troben la seua resolució cada vegada major en diferents modalitats, com les ressonàncies magnètiques; l'heterogeneïtat dels conjunts d'imatges causada, entre altres motius, per la diversitat en l'equipament emprat per a la seua adquisició; o la complexitat del seu processament i anàlisi. Diverses tecnologies clau sorgeixen de les múltiples aplicacions d'imatges mèdiques, entre elles: reconstrucció d'imatge, millora de la imatge, segmentació o diagnòstic assistit per computador.

Quant a l'evolució de les xarxes neuronals, destaquen les arquitectures que busquen crear xarxes cada vegada més profundes, com AlexNet [1], Inception Net [2] o DenseNet [3]. Les GAN –*Generative Adversarial Networks*– [4] combinen un model generador amb un model discriminador per a aprendre a generar imatges realistes, per la qual cosa s'utilitzen en reconstrucció d'imatge mèdica, millora de la qualitat d'aquestes i segmentació. Per altre costat, els mecanismes d'atenció [5] permeten la detecció automàtica de punts claus sobre la imatge. Altra tècnica important en l'ús d'aprenentatge profund amb imatge mèdica és el *transfer learning*, que permet aplicar coneixement aprés durant la resolució d'un problema per a resoldre altre problema distint. És habitual utilitzar una xarxa neuronal entrenada amb ImageNet i fer *fine tuning* sobre una tasca d'imatge mèdica per a accelerar la convergència en l'entrenament i millorar els resultats.

2.2 Estat actual de les investigacions enfocades a la detecció de COVID-19

Les investigacions sobre detecció de la malaltia COVID-19 mitjançant tècniques d'aprenentatge profund i intel·ligència artificial es troben en les seues fases inicials però en contínua evolució a causa de la novetat de la situació pandèmica i la urgència de solucions al problema mundial. El principal problema amb què es troben les primeres investigacions és la manca de dades de qualitat de pacients infectats pel virus, ja siguen imatges de radiografies, de tomografia computada

–TC, anteriorment conegudes com TAC– o resonàncies magnètiques. Per a fer front a aquest problema, la majoria d'investigacions prenen com a una primera solució utilitzar imatges corresponents a pacients amb símptomes de pneumonia anteriors a l'aparició del coronavirus.

Per als seus experiments, [6] disposa de tan sols 53 pacients amb COVID-19 en el seu conjunt de dades de 13645 pacients diferents. De la resta de pacients, 8066 no tenen pneumonia i 5526 presenten pneumonia no causada pel coronavirus. Aquest dataset és una combinació i modificació de dos repositoris d'accés obert: COVID-19 image data collection [7] i RSNA Pneumonia Detection Challenge dataset [8].

El primer d'aquests està format per imatges en format png o jpg extretes de publicacions, i és un dels principals datasets públics amb radiografies de pacients amb COVID-19. El segon dataset ha sigut creat per la Radiological Society of North America i conté imatges en format DICOM de pacients amb pneumonia.

El model proposat per [6] distingeix entre 3 prediccions: sense infecció –normal–, infecció no-COVID19 i infecció COVID-19. Els objectius d'aquestes prediccions són saber qui deu ser prioritzat per a les proves PCR i saber quina estratègia de tractament utilitzar.

Quant a la seua implementació, la xarxa proposada per [6] aplica els principis de disseny d'arquitectura residual –ResNet– i és preentrenada amb ImageNet. Utilitza l'optimitzador Adam i *learning rate scheduling*, a més de fer una bona distribució de cada tipus d'infecció a nivell de *batch*. Quant a *data augmentation*, s'hi aplica traslació, rotació, flip horitzontal i *shift* d'intensitat.

Els resultats obtinguts arriben al 92.4 % de *test accuracy*, 80 % *recall* en COVID-19 i 88.9 % *precision* en COVID-19, cosa que implica pocs casos de falsos positius de COVID-19. Els indicadors *precision* i *recall* per a normal i infecció no-COVID19 són notablement més alts que per a COVID-19. A més a més, amb l'utilització de GSInquire [9] com a mètode d'*explainability*, en la COVID-Net s'identifiquen les àrees localitzades en els pulmons que ajuden a entendre com les DNN arriben a les seues decisions.

En [10] es prova a aplicar tres xarxes CNN diferents sobre un dataset molt menut que consta de 100 imatges de rajos X de pit: 50 d'elles pertanyen a pacients amb COVID-19, les quals s'han obtingut del repositori obert de GitHub compartit pel Dr. Joseph Cohen [7], i 50 imatges normals del repositori de Kaggle "Chest X-Ray Images (Pneumonia)" [11]. Les xarxes són: ResNet50, InceptionV3 i Inception-ResNetV2. Tots els tres models proposats tenen una estructura *end-to-end* sense extracció manual de característiques i mètodes de selecció. També han sigut preentrenats amb ImageNet i utilitzen l'optimitzador Adam. Per a uns resultats més fidedignes, s'ha fet *k-fold* on k té valor 5.

Els resultats que s'obtenen en [10] són molt alts, molt probablement a causa de la poca quantitat de dades utilitzades. El model que millors resultats obté és la ResNet50, amb una *accuracy* de 98 %, *recall* igual a 96 % i *specificity* del 100 %. Per contra, els pitjors resultats s'obtenen amb la xarxa Inception-ResNetV2, quedant-se en un 87 % d'*accuracy*, 84 % de *recall* i 90 % de *specificity*.

En [12] busquen una forma de proporcionar-li a la xarxa una manera de dir "no ho sé" quan classifica les imatges amb l'objectiu d'evitar mals diagnòstics de

COVID-19. Per a açò, s'estima incertesa aproximant xarxes neuronals bayesianes –BCNN– amb *dropweights* per a millorar el rendiment del diagnòstic combinat entre humans i màquines. Basant-se en la imatge d'entrada, una xarxa pot estar segura amb una alta o baixa confiança sobre la seua decisió, indicada per la distribució predictiva posterior –*predictive posterior distribution*–.

Els autors comparen qualitativament mapes de prominència –*salience maps* [13]– produïts per diversos mètodes de l'Estat de l'Art: Class Activation Map –CAM–, Guided Backpropagation i Guided Gradient CAM. Aquests mapes serveixen per a visualitzar la incertesa, ressaltant aquelles àrees que han resultat determinants per a la classificació de les imatges.

El dataset utilitzat en [12] està format per 68 imatges de COVID, també del repositori de Github del Dr. Joseph Cohen [7], i per imatges del conjunt de Kaggle "Chest X-Ray Images (Pneumonia)" [11], de les quals 1583 són normals, 2786 corresponen a pneumonia bacterial i 1504 a pneumonia viral no-COVID. La mida de totes les imatges s'ha fixat en 224×224, utilitzant interpolació bicúbica sobre un veïnat de 4×4 píxels.

En [14] proposen un model CAAD –*Confidence-Aware Anomaly Detection*– que consisteix en un *shared feature extractor*, un mòdul de detecció d'anomalies i un mòdul de predicció de la confiança –*confidence*–. Amb aquest model busquen distingir pneumonia viral de tota la pneumonia no viral per a poder detectar els *clusters* de pneumonia viral, com la COVID-19, causats per un nou virus.

En [14] s'han utilitzat dos datasets: un per a l'entrenament amb imatges de controls sans i imatges de positius en pneumonia, i altre que també conté imatges de subjectes amb COVID-19. El primer d'aquests datasets, X-VIRAL, conté imatges en alta resolució de 390 hospitals diferents, les quals han sigut anotades entre tres radiòlegs. X-VIRAL consta de 5977 positius en pneumonia viral anteriors a la COVID-19, 18619 penumonia no viral i 18774 controls sans. El segon dataset, X-COVID, conté imatges de 106 subjectes amb COVID i 107 subjectes normals, cap d'ells utilitzat en l'entrenament, les quals provenen de 6 institucions diferents recollides durant el mes de març de 2020. Totes les imatges han sigut reescalades a 512×512. Durant l'entrenament s'aplica *data augmentation* seleccionant, de forma aleatòria, retalls de les imatges de mida 448×448; també s'aplica entre un 90 i un 110 % de zoom i es fa *flip* horitzontal.

El model CAAD, format per una xarxa de detecció d'anomalies i una xarxa de predicció de la confiança, és capaç de detectar anomalies no vistes anteriorment i depén menys de les dades etiquetades que la classificació binària. Amb aquesta proposta, els autors reformulen la detecció de pneumonia com una tasca de detecció d'anomalies basada en una classificació en una sola classe: s'assigna a cada imatge de rajos X una puntuació d'anomalia. I, a partir de la predicció de confiança, es reassignen les mostres amb baixa confiança com a mostres que tenen pneumonia però que necessiten més tests.

La xarxa de detecció d'anomalies està formada per un extractor de característiques i un mòdul de detecció d'anomalies que formen un perceptró multicapa amb tres capes ocultes de 100 neurones i una capa d'eixida d'una neurona. Com a criteri per a la confiança de la detecció d'anomalies de la xarxa de predicció de la confiança s'utilitza la funció de densitat de la probabilitat –*probability density func-*

tion, PDF–: valors alts de confiança han de pertànyer a mostres ben classificades, mentre que la confiança ha de ser baixa quan es cometen errors.

CAPÍTOL 3
Descripció del problema i de la solució proposada

Disposem d'un conjunt de dades anomenat Padchest i de la seua extensió per a Covid-19 obtinguts gràcies al Banc d'Imatge Mèdica de la Comunitat Valenciana –BIMCV–, formats per imatges de rajos X de pit. El conjunt principal, Padchest, conté imatges etiquetades segons la presència o no de pneumonia i la seua extensió afegeix imatges de pacients diagnosticats amb coronavirus, un dels símptomes del qual és la pneumonia. El BIMCV ofereix aquestes dades, degudament anonimitzades, de forma pública per a que la comunitat científica puga contribuir a la creació de solucions que ajuden al diagnòstic de la pneumonia i, sobretot, de la pneumonia provocada pel coronavirus.

En aquest projecte, aleshores, ens proposem desenvolupar models d'aprenentatge profund, o *deep learning*, capaços de detectar la presència de pneumonia sobre aquestes imatges de rajos X. També aspirem a crear models més especialitzats que sàpiguen distingir entre pneumonia comuna i penumonia provocada per coronavirus. A més, experimentarem amb Grad-CAM, una proposta per a incloure *explicabilitat* en la presentació de les prediccions de les xarxes neuronals entrenades.

3.1 CAD: *Computer Aided Diagnostics*

Els diagnòstics assistits per computador, o CAD, són procediments amb l'objectiu d'ajudar en la interpretació dels resultats de proves mèdiques. Els sistemes de CAD poden processar dades clíniques complexes i en gran quantitat tant per a inferir nous coneixements sobre aquestes dades com per a millorar el seu diagnòstic al llarg del temps. Els sistemes CAD poden ser considerats sistemes experts en medicina perquè emulen el procés de decisió que segueixen els professionals sanitaris. També poden definir-se com sistemes intel·ligents perquè utilitzen mecanismes de retroalimentació per a millorar el seu funcionament [15].

Entre els factors que han propiciat el desenvolupament de sistemes de diagnòstic per computador es troben la disponibilitat de grans quantitats de dades mèdiques rellevants per a diverses malalties i condicions i els avanços en la informàtica, especialment en els camps de la intel·ligència artificial, aprenentat-

ge automàtic i mineria de dades. Actualment, els CAD es consideren una part important del procés de diagnòstic, el qual també implica experts humans.

Per a tindre èxit, els sistemes CAD necessiten aconseguir certs objectius. El primer d'ells consisteix en tindre una notable capacitat de processament i anàlisi de grans quantitats de dades clíniques aplicant poder de computació i utilitzant algorismes especialitzats. També cal que oferisquen decisions objectives i quantitatives. A més, deuen ser efectius i eficients.

L'objectivitat dels sistemes CAD pot ajudar a realitzar diagnòstics més consistents, no influenciats per possibles experiències o prejudicis dels professionals sanitaris [16] ni afectats negativament per la fatiga que aquests poden sofrir a causa de les cansades jornades laborals [17]. A més, aquestos sistemes poden ajudar a disminuir la bretxa entre professionals amb molta experiència i professionals amb poca experiència en relació a diagnosi mèdica.

Quant a l'eficàcia i l'eficiència del sistemes CAD, es fa referència, especialment, a la detecció de les malalties en les etapes primerenques, quan solen ser asimptomàtiques i més fàcilment tractables. La detecció primerenca de les malalties és crucial per a poder aplicar un tractament adequat com menys perjudicial per al pacient es puga, tant per a evitar la progressió de la malaltia com procediments invasius com poden ser les cirurgies. A més, els sistemes de CAD poden ajudar a compensar la capacitat limitada dels humans per a trobar anomalies, inclús quan s'utilitza equipament especialitzat. Els sistemes CAD poden, per exemple, detectar microcalcificacions que els professionals sanitaris no havien sabut reconéixer [18]. També poden reduir el temps que es tarda en trobar calcificacions poc clares [19].

Els sistemes CAD es poden diferenciar segons el tipus de dades que utilitzen: sons i senyals de diversos òrgans, imatges del cos humà o dades que s'han obtingut a partir de tests de laboratori. Es poden combinar els tipus de dades utilitzades segons la detecció de cada malaltia així ho necessite.

3.2 Mètodes de Deep Learning

3.2.1. Xarxes Neuronals Convolucionals

Per al problema abordat en aquest projecte, consistent en la classificació d'imatges de rajos X segons si mostren símptomes de pneumonia, ja siga causada pel coronavirus o no, s'ha decidit utilitzar xarxes neuronals convolucionals. Aquest tipus de xarxes solen ser aplicades en l'anàlisi d'imatges visuals, o visió per computador.

Les xarxes neuronals convolucionals troben patrons a la imatge i els assignen pesos. A més, són capaces de capturar amb bons resultats aquelles dependències espacials que apareixen a la imatge a partir de l'aplicació de filtres rellevants. El seu objectiu consisteix a reduir les imatges fent que aquestes siguen més fàcils de processar però sense perdre aquelles qualitats o característiques determinants per a extraure una bona predicció. Aquesta característica és important per a obtindre arquitectures escalables a conjunts de dades de mides considerables.

El component més destacable d'una xarxa neuronal convolucional són les convolucions. Aquestes operacions són realitzades per un filtre, anomenat kernel, sobre una part de la imatge multiplicant els elements entre sí. La mida d'aquest filtre determina sobre quants elements es realitza la convolució, i el desplaçament del kernel sobre la matriu de la imatge el marcarà el *stride*. Seguint aquests passos, la convolució permet centrar l'atenció sobre zones de la imatge que ajuden a diferenciar entre les classes possibles.

Els models emprats en aquest projecte estan formats per la repetició de blocs de capes composats per capes convolucionals amb diverses mides de kernel, capes *batch normalization* i capes *max-pooling*. Com que el problema a resoldre consisteix en la classificació d'imatges, l'eixida de l'última capa serà sempre donada per una capa amb activació *softmax*. En els models utilitzats, s'experimenta amb la repetició d'una mateixa estructura convolucional en una xarxa augmentant progressivament el número de filtres. També s'experimenta amb la mida de les primeres convolucions aplicades sobre les imatges.

3.2.2. GradCAM

La tècnica Grad-CAM –*Gradient-weighted Class Activation Mapping*– [20] és una proposta per a generar "explicacions visuals" sobre models de xarxes neuronals convolucionals, buscant millorar l'*explicabilitat* d'aquests. Grad-CAM utilitza els gradients que arriben a l'última capa convolucional per a produir un mapa de localització aproximat que ressalta les regions importants de la imatge per predir la classe d'una mostra.

Un avantatge de Grad-CAM en comparació amb altres tècniques anteriors que busquen aquesta *explicabilitat* és la seua capacitat d'aplicar-se a una àmplia varietat de models convolucionals sense necessitar modificar l'arquitectura dels models ni reentrenar-los. La tècnica més rellevant en comparació a Grad-CAM, és CAM –*Class Activation Mapping*–[21], la qual modifica les arquitectures de xarxes convolucionals de classificació reemplaçant les capes *fully-connected* per capes convolucionals i *global average pooling*. Aquesta modificació fa que CAM sols puga ser aplicable a una selecció de xarxes, deixant fora d'aquesta selecció tasques com subtitulació d'imatge –*image captioning*– o *Visual Question Answering*.

3.3 Descripció de les ferramentes software i hardware utilitzades

3.3.1. Tensorflow

TensorFlow és un sistema d'aprenentatge automàtic que opera a gran escala i en entorns heterogenis. TensorFlow utilitza gràfics de flux de dades per representar la computació, l'estat compartit i les operacions que muten aquest estat. Mapifica els nodes d'un gràfic de flux de dades a través de moltes màquines d'un mateix clúster, i dins d'una màquina a través de múltiples dispositius informàtics, incloent CPU, GPU i TPU [22, 23].

TensorFlow és compatible amb una gran varietat d'aplicacions, enfocat en l'entrenament i la inferència en xarxes neuronals profundes. TensorFlow va ser llançat com a projecte de codi obert per Google, diversos serveis del qual l'empren, i s'utilitza àmpliament per a la recerca d'aprenentatge automàtic.

3.3.2. Keras

Per construir i entrenar models amb TensorFlow, la millor opció, i la més fàcil, és utilitzar l'API d'alt nivell Keras. Aquesta API no només és simple i consistent, sinó que també redueix el nombre d'accions requerides a l'usuari per implementar els casos d'ús més comuns. Gràcies a aquestes característiques, Keras se centra en permetre una ràpida experimentació. Keras està escrit en Python i pot executar per sobre de TensorFlow, CNTK o Theano. A més, admet tant xarxes convolucionals com recurrents i també combinacions d'aquestes [24].

Keras és el *framework* d'aprenentatge profund més utilitzat tant a la indústria com a la comunitat investigadora sols per darrere del propi TensorFlow.

3.3.3. GPU

Les GPU utilitzades han estat disponibles per a aquest projecte gràcies al centre d'investigació Pattern Recognition and Human Language Technology –PRHLT–. S'ha disposat de diverses NVIDIA GeForce RTX 2080 i també NVIDIA GeForce GTX 1080.

Taula 3.1: Comparació de les GPU

	RTX 2080	GTX 1080
CUDA cores	2944	2560
Memòria estàndard	8 GB	8 GB

Cal destacar la importància de poder utilitzar GPU en lloc de CPU, especialment per a treballar amb imatges, gràcies a la seua velocitat de processament.

CAPÍTOL 4
Conjunt de dades

Per a aquest projecte s'han utilitzat les dades proporcionades pel Banc D'Imatge Mèdica de la Comunitat Valenciana, BIMCV.

El conjunt de dades PadChest –*Pathology Detection in Chest Radiographs*– és un dels datasets públics de radiografies de pit més complets, tant per quantitat d'imatges com per l'etiquetat exhaustiu d'aquestes. A més, és l'únic que conté extractes dels informes font escrits en espanyol. Les etiquetes de les imatges estan *mapejades* a la terminologia estàndard del *Unified Medical Language System* –UMLS– i, per tant, es poden utilitzar sense que l'idioma original supose cap problema. A més, aquestes variables s'identifiquen amb etiquetes anatòmiques, les imatges es guarden en alta resolució i s'inclou també informació molt extensa que inclou la demografia del pacient, el tipus de projecció i els paràmetres d'adquisició, entre d'altres.

L'etiquetat d'una part del conjunt d'imatges per a crear el *ground-truth* de referència s'ha fet de forma manual per metges formats per a maximitzar la fiabilitat d'aquest *ground-truth*. A partir d'aquest etiquetat manual, s'ha realitzat un etiquetat automàtic basat en Deep Learning extraient la informació dels informes restants.

Aquest conjunt de dades està format per radiografies de pit, les quals han sigut interpretades i descrites en informes per 18 radiòlegs de l'Hospital Universitari de Sant Joan d'Alacant, des de gener de 2009 fins a desembre de 2017.

El conjunt de dades PadChest està composat per un total de 160868 imatges corresponents a 67625 pacients. D'aquestes més de 160000 imatges, 39039 han sigut etiquetades manualment per metges formats. La resta de les imatges han sigut etiquetades de forma automàtica utilitzant mètodes supervisats basats en xarxes neuronals recurrents amb mecanismes d'atenció. Per aquest motiu i ja que el volum de dades en l'extensió amb imatges de pacients amb COVID-19 és molt menor, s'han seleccionat per als experiments sols imatges anotades manualment. D'entre les 39039 imatges etiquetades manualment, sols 11989 corresponen a escàners de l'estàndard "*Postero-Anterior*". Tan sols s'han seleccionat imatges que segueixen aquest estàndard d'adquisició perquè és l'utilitzat per a detectar COVID-19.

4.0.1. Preprocessament

El conjunt Padchest conté imatges classificades en quatre grups: Control, Pneumonia, Pneumonia Infiltrated i Infiltrated. D'aquestes imatges sols hem utilitzat aquelles que pertanyen als grups Control i Pneumonia, ja que les imatges dels altres dos grups han sigut classificades segons una etiqueta més incerta respecte a la patologia que presenten. Aquestes imatges les hem combinat amb les imatges de pacients infectats per coronavirus de l'extensió BIMCV-COVID19. D'aquest segon conjunt, s'ha fet una selecció, ja que les imatges s'han pres des de diferents angles, o projeccions. S'han eliminat aquelles imatges que no corresponguen a les projeccions *anteroposterior* o *posteroanterior*, ja que aquestes són les projeccions utilitzades en Padchest.

Per al seu preprocessament s'ha tingut en ment l'homogeneïtzació del conjunt de dades que s'ha obtingut de la combinació dels dos conjunts mencionats. Buscant l'eliminació de les diferències en les imatges quant a escala de valors o mida ens assegurem de que els models aprenguen a classificar segons el contingut de la imatge i no segons el seu format. Com que les imatges de BIMCV-COVID19 són més recents que les de Padchest, podia donar-se el cas que les imatges de cada conjunt tingueren certes característiques diferenciadores causades, per exemple, per l'ús d'una maquinària diferent per a la seua captació.

El preprocessament ha sigut el següent:

- en cas de que la imatge tinga el color invertit (MONOCHROME1), s'inverteix la imatge
- es canvia la mida de la imatge a 524 x 524
- es normalitzen els valors dels píxels entre 0 i 255

Per a fer la divisió de les imatges en conjunts de training, validació i test, s'ha escollit una divisió 60-20-20. Açò vol dir que un 60 % de les imatges es destinen a training i la resta es divideix entre validació i test. Però, com que les classes estan desbalancejades, és convenient aplicar alguna tècnica per a evitar la inclinació dels models cap a la classe majoritària.

La solució escollida consisteix a assignar el mateix nombre de mostres de cada classe tant en validació com en el test i deixar tota la resta per a training però, durant l'entrenament, utilitzar el mateix nombre de mostres de cada classe en cada *batch*, les quals seran seleccionades aleatòriament pel *DataGenerator*. Així, el model en tot moment veurà la mateixa proporció de mostres de cada classe. D'aquesta manera, s'aprofiten totes les mostres de les quals es disposa, sense necessitat d'eliminar part del dataset per a aconseguir aquest equilibri de classes.

CAPÍTOL 5
Experimentació

Per a l'experimentació, s'han dissenyat 4 topologies principals de xarxes neuronals convolucionals, cadascuna d'aquestes amb diverses variants diferenciades per la mida del kernel aplicat a la imatge d'entrada. Per altra banda, les diferències entre cadascuna de les 4 topologies principals es troben en el tipus de bloc utilitzat per a realitzar les reduccions graduals i en les capes que segueixen a estos blocs.

S'han utilitzat aquests models per a fer una classificació en les 3 classes: Control, Pneumonia i Covid-19. També s'han provat com a classificadors binaris amb les següents combinacions:

- Control vs. Pneumonia
- Control vs. Covid-19
- Pneumonia vs. Covid-19
- Control vs. (Pneumonia + Covid-19)

5.1 Classificació en cascada

A més a més, s'ha plantejat el disseny de classificadors "en cascada", formats per dos classificadors binaris: Control vs. (Pneumonia + Covid-19) i Pneumonia vs. Covid-19. Aquestos dos classificadors s'entrenen per separat, cadascun amb les dades corresponents, i el funcionament amb les dades de test és el següent:

- es classifiquen les imatges utilitzant el classificador Control vs. (Pneumonia + Covid-19)
- aquelles imatges que han sigut etiquetades com a (Pneumonia + Covid-19) es classifiquen ara utilitzant el classificador Pneumonia vs. Covid

Aquesta proposta, aleshores, consisteix en detectar primer si existeix malaltia a les imatges i, a continuació, decidir si les imatges identificades com a infectades mostren evidències de pneumonia comú o de pneumonia provocada per Covid-19. Aquest disseny permet utilitzar topologies diferents en cada fase de la

classificació, atenent a la possibilitat que el primer classificador se centre en detectar anomalies que indiquen malaltia i el segon classificador estiga especialitzat en identificar les diferències entre els dos tipus de pneumonia.

5.2 Blocs convolucionals

Els models creats es basen en la repetició, modificant cada vegada el nombre de filtres, d'un bloc format per tres capes convolucionals. Cada bloc s'utilitza en diversos models, diferenciats per les convolucions inicials o per les operacions aplicades en acabar l'últim bloc. La creació d'aquestes xarxes s'ha fet de forma progressiva i per això la seua nomenclatura és numèrica, marcant les variacions alfabèticament. A continuació s'expliquen els blocs utilitzats i les xarxes on s'utilitzen. Els esquemes complets de les xarxes s'inclouen en l'Apèndix.

5.3 Models 4

El bloc bàsic del primer model dissenyat està format per tres capes Convolucionals en 2D seguides, cadascuna d'elles, per una capa d'Activació ReLU i, finalment, una capa MaxPooling2D. La primera i tercera capes Convolucionals apliquen convolucions de mida 1×1, mentre que la segona aplica convolucions 3×3. La capa MaxPooling2D aplica un *pool* de mida 2×2. Aquest bloc, representat en la Figura 5.1, es repeteix 6 vegades en cada model: les dues primeres vegades les capes convolucionals utilitzen 128 filtres; les tres vegades següents, 256; i l'última, 512. A continuació d'aquesta última repetició del bloc, s'utilitza una capa Convolucional 2D amb convolució 1×1 i 512 filtres, seguida d'una capa GlobalMaxPooling2D i dos Fully Connected. L'output s'aconsegueix amb activació *softmax*.

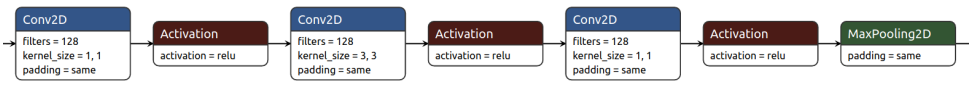


Figura 5.1: Bloc del model 4

Per a aquest primer model, s'han creat 4 variacions –4A, 4B, 4C i 4D, representades a les Figures A.1, A.2, A.3 i A.4 de l'Apèndix– que es diferencien en les capes d'entrada. El model 4A utilitza com a capes d'entrada 3 capes Convolucionals 2D, amb mida de convolucions 7×7, 9×9 i 11×11, respectivament, les quals concatena abans de seguir amb el bloc bàsic. El model 4B utilitza 2 capes Convolucionals 2D com a entrada, amb convolucions 5×5 i 7×7, les quals també concatena. El model 4C segueix aquest mateix patró, però amb convolucions 7×7 i 9×9. I el model 4D aplica convolucions 9×9 i 11×11.

5.4 Models 5

El bloc bàsic del segon model, representat a la Figura 5.2, és igual que el del primer model però eliminant la capa d'activació ReLU entre la primera i segona capes convolucionals. En totes les variacions també es repeteix 6 vegades amb la mateixa combinació de filtres que el model anterior i acaba seguit d'una capa Convolucional 2D amb convolució 1×1.

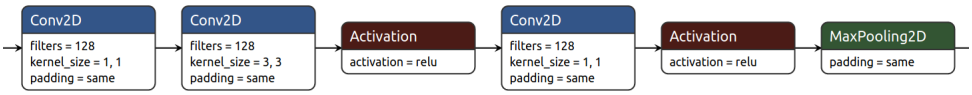


Figura 5.2: Bloc del model 5

Per a aquest model, s'han creat 3 variacions: 5A, 5B i 5C. La primera variació, la qual es troba representada a la Figura A.5, consta de dues capes Convolucionals 2D com a capes d'entrada, una amb convolucions 7×7 i l'altra amb convolucions 9×9, ambdues seguides pel bloc bàsic i, a continuació, una capa GlobalMaxPooling2D. Les dues capes GlobalMaxPooling2D es concatenen i, finalment, el model acaba amb dues capes Fully Connected.

Les altres dues variacions, a les Figures A.6 i A.7, són més simples, amb una sola capa Convolucional 2D com a capa d'entrada. Aquesta capa també és seguida pel bloc bàsic. Després del bloc, es fa un Flatten a la seua eixida i aquesta es passa per una capa Fully Connected que dona el resultat de la xarxa amb activació *softmax*. Una d'aquestes variacions utilitza convolucions 7×7 per a la primera capa i l'altra utilitza convolucions 9×9.

5.5 Models 6

Per als models 6A s'ha utilitzat un bloc que realitza convolucions de mides 3×3 i 5×5. Després de cada capa convolucional, s'uneix una capa d'activació ReLU. El bloc utilitzat en el model 6B segueix aquesta mateixa estructura però afegeix una capa BatchNormalization després de cada capa d'activació ReLU. Aquests blocs es mostren a les Figures 5.3 i 5.4.

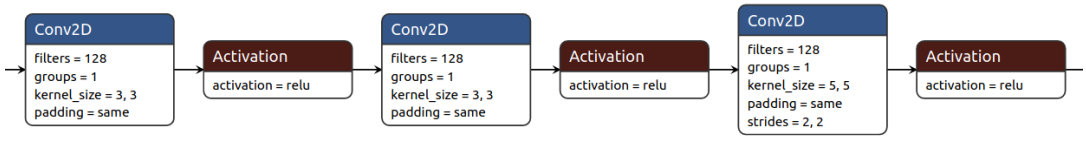


Figura 5.3: Bloc del model 6A

El model 6A adopta una estructura similar al model 4A pel fet d'utilitzar 3 capes Convolucionals 2D com a entrada, amb convolucions 7×7, 9×9 i 11×11. No obstant, en aquest cas s'aplica el bloc convolucional de la Figura 5.3 abans de fer la concatenació de les eixides de les capes. El bloc convolucional es repeteix sis vegades: les tres primeres amb 128 filtres i les 3 següents amb 256

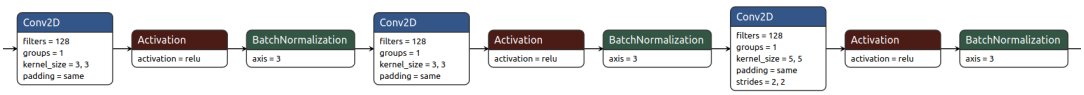


Figura 5.4: Bloc del model 6B

filtres. Després de la capa de concatenació, es fa un Flatten i l'eixida és donada per una capa Fully Connected amb activació *softmax*.

El model 6B realitza convolucions de mida 7×7 a la imatge d'entrada i a continuació aplica el bloc de la Figura 5.4 sis vegades amb la mateixa combinació de filtres que el model 6A. Les dues últimes capes són les mateixes també que el model 6A: una capa Flatten després de l'última repetició del bloc convolucional i una capa Fully Connected amb eixida *softmax*.

Quant al model 6C, el bloc convolucional utilitzat és idèntic al dels models 4, tal i com es pot veure a la Figura 5.5, però afegeix BatchNormalization abans de les activacions ReLU. En aquest model, la imatge passa primer per una capa convolucional amb convolucions de mida 7×7 i després per una altra amb mida 5×5. A continuació, s'inclou el bloc descrit també sis vegades, però amb les següents quantitats de filtres: 100, 150, 200, 250, 300 i 350. L'esquema del model 6C es pot veure a la Figura A.10.

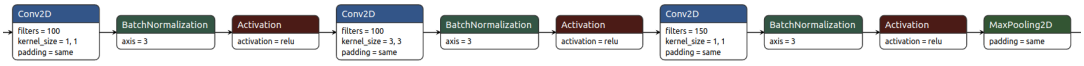


Figura 5.5: Bloc del model 6C

5.6 Models 7

Per als models 7 s'han utilitzat dos blocs convolucionals diferents. El primer d'aquests blocs, il·lustrat a la Figura 5.6, és igual al bloc 6A, però s'ha optat per definir el *padding* com a *valid* en lloc de *same*. El segon bloc, mostrat a la Figura 5.7, imita el bloc 6C, també modificant el tipus de *padding* a *valid*.

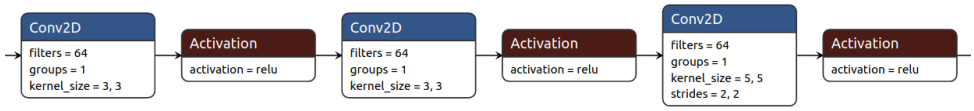


Figura 5.6: Bloc dels models 7A i 7D

El model 7A, representat a la Figura A.11, realitza convolucions de mida 5×5 amb *padding valid* sobre la imatge d'entrada i després utilitza el bloc 7A quatre vegades amb 64, 128, 128 i 256 filtres. Seguidament, s'han afegit dues capes convolucionals més també amb *padding valid*, esta vegada amb mida de convolució 3×3, una capa GlobalMaxPooling2D i altra capa Fully Connected amb activació *softmax*.

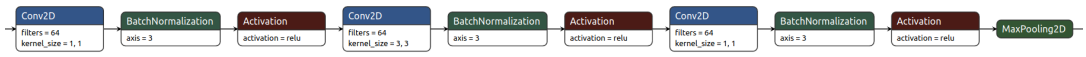


Figura 5.7: Bloc dels models 7B i 7C

La primera capa dels models 7B i 7C és una capa convolucional; en el model 7B les convolucions tenen dimensió 7×7 i en el model 7C, 9×9. Després s'empra el bloc 7B cinc vegades amb les següets quantitats de filtres: 64, 128, 128, 256 i 512. Finalment, les xarxes acaben amb una capa GlobalMaxPooling2D i dues Fully Connected, la primera amb activació ReLU i la següent amb activació *softmax*. Aquestos dos models es poden observar a les Figures A.13 i A.14.

El model 7D té la mateixa estructura que el model 7A, canviant la primera capa convolucional de la xarxa per una amb mida de convolució 11×11 amb *strides* de dimensió 3×3 i eliminant l'última capa convolucional. Aquest model es mostra a la Figura A.12.

5.7 Data augmentation

Una revisió de les imatges ha evidenciat les desigualtats entre aquestes. Aquestes diferències provenen del moment de l'adquisició de la radiografia: canvis en la postura del pacient, enquadrament de la imatge, diferent rang d'intensitat, etc. A més, les diferències són, generalment, més exagerades en les imatges de pacients amb coronavirus, ja que moltes es van realitzar de forma menys acurada a com es fa normalment per poder donar abast. Es per això que s'ha volgut fer ús de tècniques de *data augmentation* que asseguren que les xarxes dissenyades puguen adaptar-se a aquestes irregularitats.

Durant l'entrenament, s'ha aplicat *data augmentation* a, aproximadament, un 30 % de les imatges. A aquestes imatges se'ls ha aplicat, de forma aleatòria, una o més de les tècniques següents:

- Rotació: es gira la imatge fins a 15º en qualsevol dels dos sentits.
- Desplaçament: es desplaça la imatge fins a un màxim 20 píxels tant en un eix com en l'altre, sent aquest desplaçament no necessàriament igual en els dos eixos.
- Zoom: s'amplia o es redueix el contingut de la imatge, mantenint la mida original, fins un 10 %.
- Variació de la intensitat: es multipliquen els valors de l'array de la imatge per un factor entre 0,8 i 1,2.

CAPÍTOL 6
Resultats

A continuació es comenten els resultats obtinguts més destacables per a cada combinació d'etiquetes. L'avaluació dels models es fa a partir de la mesura de l'*accuracy*, *precision*, *recall*, *f1-score* i àrea sota la corba ROC –*receiver operating characteristic*– o AUC en aquelles classificacions en dos grups.

Per a definir aquestes mètriques cal tindre en compte els conceptes de *true positives* –TP, mostres classificades en una classe que sí pertanyen a eixa classe–, *false positives* –FP, mostres classificades en una classe que no pertanyen a eixa classe–, *true negatives* –TN, mostres no classificades en una classe que no pertanyen a eixa classe– i *false negatives* –FN, mostres no classificades en una classe però que sí pertanyen a eixa classe–.

L'*accuracy* correspon a un recompte d'aquelles mostres que han sigut correctament classificades i es calcula com la proporció de resultats correctes sobre el total de casos examinats:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

L'*accuracy* és un valor global per a tota la classificació, és a dir, no canviarà encara que es calcule sobre cadascuna de les classes. Per contra, *precision*, *recall* i *f1-score* sí són dependents de la classe per a la qual es calculen. La combinació d'aquests valors indica com de correcta ha sigut la distinció entre les classes.

Calculant *precision* se sap la proporció de mostres que realment pertanyen a un grup d'entre totes les que han sigut classificades en aquest grup:

$$Precision = \frac{TP}{TP + FP}$$

Amb *recall* es calcula la quantitat de mostres pertanyents a una classe que s'han arribat a classificar en eixa classe:

$$Recall = \frac{TP}{TP + FN}$$

Quant a *f1-score*, aquesta mesura proporciona una mitjana harmònica de *precision* i *recall* que permet avaluar les dues mètriques anteriors en conjunt:

$$Fscore = 2 \times \frac{precision \times recall}{precision + recall}$$

Finalment, la corba ROC és la representació de la ràtio de vertaders positius –TP– front a la ràtio de falsos positius –FP– segons varia el llindar de discriminació o valor a partir del qual considerem un cas com positiu. Per aquest motiu, la corba ROC sols la calculem en aquelles classificacions binàries, on es distingeix entre únicament dues classes. El seu anàlisi ajuda a discriminar models possiblement òptims d'aquells que no ho són tant. Com a avantatge de la corba ROC, cal dir que és independent de la distribució de les classes en la població o, com que parlem de diagnòstic, independent de la prevalença d'una malaltia en la població.

6.1 Classificació en dues classes

6.1.1. Control vs. Pneumonia

Amb els classificadors entrenats per a distingir entre imatges Control i imatges amb símptomes de Pneumonia, s'aconsegueixen resultats propers al 80 % d'*accuracy*. Els millors resultats quant a *accuracy* s'aconsegueixen amb el model 7A, arribant al 79,6 %. Aquest model també és el que trau millor resultats de *recall* i *f1-score* per a la classe Pneumonia, amb un 0,79 per a ambdues mètriques.

Entre els pitjors resultats estan aquells del model 6B que, com es pot comprovar per l'alt *recall* de la classe Control i la baixa *precision* d'aquesta mateixa classe combinats amb el baix *recall* de la classe Pneumonia, tendeix a classificar la majoria de mostres en la classe Control.

Els resultats que s'obtenen amb alguns dels altres models es poden veure en la Taula 6.1

Taula 6.1: Resultats Control vs. Pneumonia

Model	Acc.	Precision		Recall		F1-score	
		Control	Pneum.	Control	Pneum.	Control	Pneum.
4A	75,4 %	0,71	0,81	0,85	0,66	0,77	0,73
4B	**77,7 %**	0,74	0,82	0,85	0,71	0,79	0,76
4C	76,8 %	0,75	0,79	0,80	0,73	0,77	0,76
5A	**77,6 %**	0,75	0,81	0,83	0,73	0,79	0,76
5B	76,0 %	0,73	0,80	0,82	0,70	0,77	0,75
6B	53,1 %	0,52	0,71	0,96	0,11	0,67	0,19
7A	**79,6 %**	0,79	0,80	0,80	0,79	0,80	0,79
7C	68,6 %	0,69	0,68	0,67	0,70	0,68	0,69

A la Figura 6.1 es mostra la representació de la corba ROC per al model 7A. L'àrea que crea aquest model, que és el que ha aconseguit una major *accuracy*, és de 0,86.

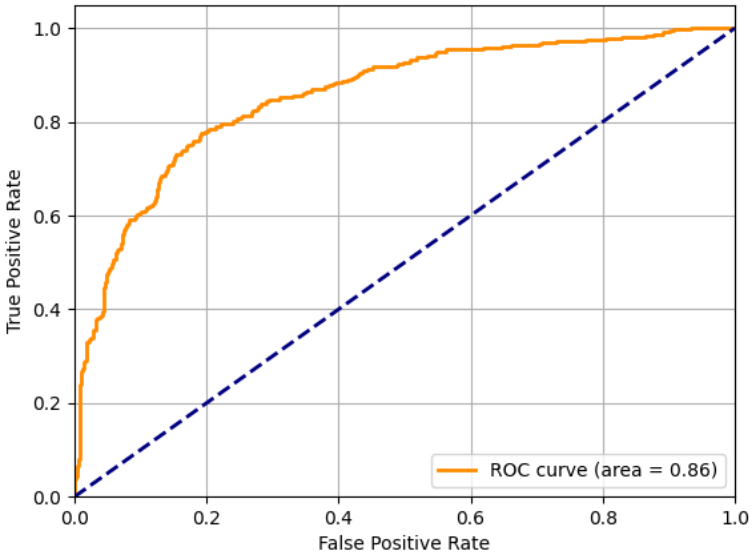


Figura 6.1: Representació de la corba ROC del model 7A per a les classes Control i Pneumonia.

6.1.2. Control vs. Covid-19

Els resultats de la classificació entre Control i Covid-19 són més alts que per a Control i Pneumonia, com es pot veure a la Taula 6.2. S'aconsegueix una *precision* molt propera a l'1 per a la classe Covid-19 en quasi tots els models, la qual cosa indica pocs falsos positius. El *recall* es queda sobre un 0,8, la qual cosa significa una alta taxa de detecció de la malaltia.

El model amb menys falsos negatius és el model 6A, que obté un 0,86 de *recall* en Covid-19, però la seua *precision* es queda en 0,68, la major taxa de falsos positius d'entre els models analitzats.

Taula 6.2: Resultats Control vs. Covid-19

Model	Acc.	Precision		Recall		F1-score	
		Control	Covid-19	Control	Covid-19	Control	Covid-19
4A	87,7 %	0,81	0,99	0,99	0,75	0,90	0,85
4B	88,7 %	0,83	0,98	0,99	0,78	0,90	0,87
4C	**90,0 %**	0,85	0,98	0,99	0,80	0,91	0,88
4D	**89,4 %**	0,84	0,99	0,99	0,78	0,91	0,87
5A	83,2 %	0,76	0,98	0,99	0,66	0,86	0,79
5B	**89,0 %**	0,85	0,94	0,96	0,82	0,90	0,88
6A	74,5 %	0,83	0,68	0,64	0,86	0,73	0,76
6C	69,7 %	0,70	0,70	0,75	0,64	0,72	0,67
7A	76,7 %	0,70	0,96	0,98	0,53	0,82	0,68
7D	75,8 %	0,69	0,96	0,98	0,51	0,81	0,67

L'àrea sota la corba ROC generada pel model 4C, la qual es mostra a la Figura 6.2, és de 0,98.

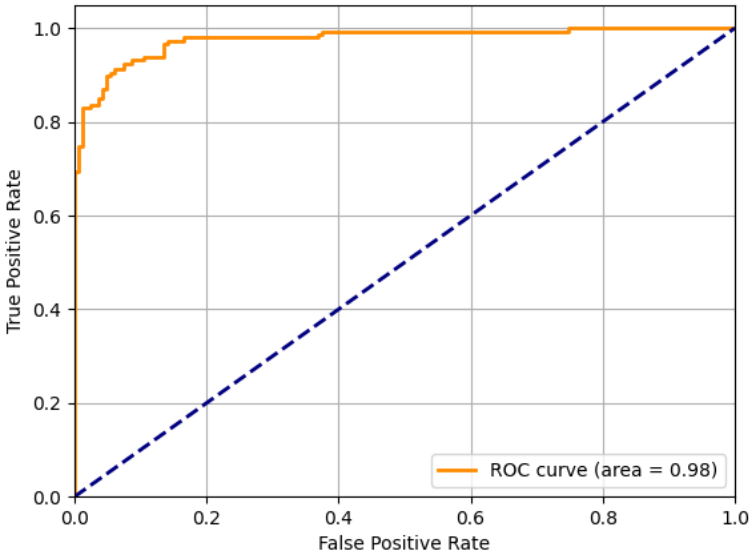


Figura 6.2: Representació de la corba ROC del model 4C per a les classes Control i Covid-19.

6.1.3. Covid-19 vs. Pneumonia

Els resultats de la classificació entre Covid-19 i Pneumonia es poden veure a la Taula 6.3.

Resulten interessants els valors d'*accuracy* aconseguits en comparació amb aquells de les classificacions anteriors. En general, s'ha aconseguit una major *accuracy*, que ronda el 90 %. La *precision* per a la classe Covid-19 és lleugerament més alta, superant el 0,9 en quasi tots els models.

Taula 6.3: Resultats Covid-19 vs. Pneumonia

Model	Acc.	Precision		Recall		F1-score	
		Covid-19	Pneum	Covid-19	Pneum.	Covid-19	Pneum.
4A	**90,0 %**	0,95	0,86	0,83	0,96	0,89	0,91
4B	**91,6 %**	0,96	0,88	0,86	0,97	0,91	0,92
4C	89,4 %	0,93	0,87	0,84	0,94	0,88	0,90
4D	86,5 %	0,93	0,82	0,77	0,95	0,84	0,88
5A	**92,6 %**	0,91	0,94	0,94	0,91	0,92	0,93
5B	89,4 %	0,97	0,84	0,80	0,98	0,88	0,91
5C	**94,2 %**	0,95	0,93	0,93	0,96	0,94	0,95
6B	79,4 %	0,72	0,92	0,94	0,66	0,81	0,77
7B	61,9 %	0,97	0,58	0,20	0,99	0,34	0,73
7D	69,4 %	0,89	0,64	0,40	0,96	0,55	0,77

L'àrea sota la corba ROC per al model 5C és de 0,98, tal i com es pot veure en 6.3.

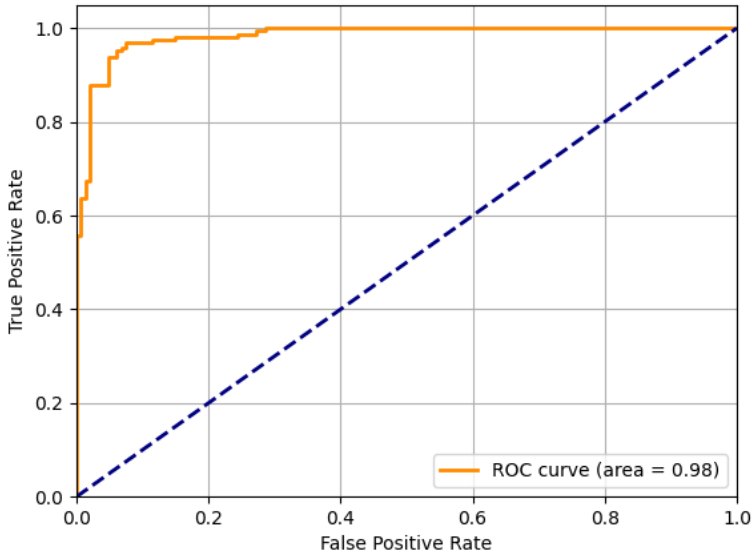


Figura 6.3: Representació de la corba ROC del model 5C per a les classes Pneumonia i Covid-19

6.1.4. Control vs. (Pneumonia + Covid-19)

Quant als resultats obtinguts en la classificació entre Control i els dos tipus de pneumonia (Pneumonia i Covid-19) en conjunt, veiem que s'arriba fins a un 83,5 % d'*accuracy* amb el model 7B. Els següents valors més alts en *accuracy* els trauen els models 4B –80,8 %–, 5C –80,7 %– i 7D –80,7 %–. D'entre aquests, el model que més redueix el nombre de falsos negatius quant a detecció de pneumonia és el model 4B, ja que aconsegueix un *recall* igual a 0,78, superior al de la resta.

Taula 6.4: Resultats Control vs. (Pneumonia + Covid-19)

Model	Acc.	Precision		Recall		F1-score	
		C	(P + C-19)	C	(P + C-19)	C	(P + C-19)
4A	78,9 %	0,77	0,82	0,85	0,73	0,81	0,77
4B	**80,8 %**	0,80	0,82	0,83	0,78	0,82	0,80
4C	77,5 %	0,73	0,84	0,88	0,66	0,80	0,74
5A	78,9 %	0,76	0,83	0,86	0,71	0,81	0,77
5B	80,2 %	0,76	0,86	0,89	0,71	0,82	0,78
5C	**80,7 %**	0,76	0,88	0,91	0,70	0,83	0,78
6C	63,1 %	0,62	0,64	0,70	0,56	0,66	0,60
7A	77,4 %	0,73	0,85	0,89	0,65	0,80	0,74
7B	**83,5 %**	0,79	0,89	0,91	0,75	0,85	0,82
7D	**80,7 %**	0,78	0,84	0,87	0,75	0,82	0,79

6.2 Classificació en tres classes

Per a la classificació en tres classes –Control, Covid-19 i Pneumonia– s'ha optat per comparar dos aproximacions diferents. La primera d'elles consisteix a entrenar classificadors per a distingir entre les tres classes. La segona consisteix en combinar dos classificadors binaris, el primer, per a detectar la presencia o no de malaltia i, el segon, per a distingir entre quin dels dos tipus de malaltia es troba present a cada imatge. Per a la primera part d'aquesta classificació s'han combinat les mostres de Covid-19 i Pneumonia en una sola classe, aquella que mostra símptomes de malaltia. Les mostres que siguen classificades com a pertanyents a aquest conjunt passen a ser classificades per un segon model, que distingeix entre Covid-19 i Pneumonia.

6.2.1. Utilitzant un sol classificador

Com podem veure a la Taula 6.5, alguns models presenten una *accuracy* molt baixa, inferior al 50 %. En el cas dels models 6C i 7B, s'ignora completament la classe Covid-19. Els models 6B i 7C obtenen una *accuracy* inclús inferior al 40 %, però en aquest cas sí classifiquen algunes imatges com a Covid-19. Però, si ens fixem en els valors de *recall* d'aquests dos models per a cada classe –0,70 en Control i 0,21 en Covid-19 i Pneumonia per al model 6B; 0,95 en Control, 0,04 en Covid-19 i 0,13 en Pneumonia per al model 7C–, descobrim que aquests models tendeixen a classificar la majoria de les mostres en la mateixa classe: Control.

Quant als models amb millors resultats, observem que el model 4C aconsegueix un 80 % d'*accuracy*. Aquest model segueix tenint preferència per la classe Control, segons els resultats de *precision* –0,58– i de *recall* –0,91–; i sembla tindre més problemes amb la classe Pneumonia, ja que es queda amb un 0,48 de *recall*, a diferència del 0,74 que aconsegueix en la classe Covid-19. El següent model amb millor *accuracy* és el 4B que, tot i que es queda en un 73,3 % d'*accuracy*, millora els resultats de *f1-score* de dues de les tres classes en comparació amb el model 4D.

Taula 6.5: Resultats Control vs. Pneumonia vs. Covid-19

Model	Acc.	Precision			Recall			F1-score		
		C	C-19	P	C	C-19	P	C	C-19	P
4A	71,8 %	0,61	0,96	0,72	0,83	0,65	0,67	0,79	0,78	0,69
4B	**73,3 %**	0,62	0,94	0,73	0,82	0,78	0,60	0,71	0,85	0,66
4C	**80,0 %**	0,58	0,89	0,85	0,91	0,74	0,48	0,71	0,81	0,61
4D	70,8 %	0,58	0,94	0,78	0,88	0,68	0,55	0,70	0,79	0,65
5A	70,1 %	0,64	0,97	0,64	0,77	0,61	0,71	0,70	0,75	0,67
5C	**72,2 %**	0,64	0,98	0,68	0,80	0,65	0,71	0,71	0,78	0,69
6B	37,9 %	0,34	0,43	0,49	0,70	0,21	0,21	0,46	0,28	0,29
6C	43,4 %	0,44	0,00	0,43	0,82	0,00	0,44	0,57	0,00	0,43
7A	67,6 %	0,57	0,96	0,68	0,88	0,59	0,55	0,69	0,73	0,61
7B	41,9 %	0,40	0,00	0,49	0,91	0,00	0,30	0,56	0,00	0,37
7C	38,8 %	0,37	0,43	0,63	0,95	0,04	0,13	0,53	0,07	0,22

6.2.2. Combinant dos classificadors

S'ha buscat combinar classificadors de forma estratègica per a millorar els resultats.

La primera combinació consisteix en l'ús del mateix classificador, el 5C, per a les dos fases de la classificació. Si observem en la Taula 6.5, el model 5C és un dels dos models amb major *precision* per a la classe Control, és a dir, un dels models que menys mostres "malaltes" classifica com a "sanes". Quant a la seua eficàcia distingint entre les classes Covid-19 i Pneumonia, veiem a la Taula 6.1, que és el model amb major *accuracy*. Finalment, el resultat d'aquesta combinació de models mostra una millora de quasi un 7 % d'*accuracy* respecte a l'ús del model 5C per a classificar directament en tres classes. També millora la *precision* en Control i Pneumonia i el *recall* en Control i Covid-19, però baixa el *recall* en Pneumonia.

Per a la resta de combinacions de dos models iguals s'ha seguit el mateix criteri d'elecció: escollir els models amb major *accuracy* i *precision* més alta per a la classe Control en la classificació en tres classes. Quant a la combinació de dos models diferents, s'han escollit per al segon classificador aquells models amb millors resultats en la classificació entre Pneumonia i Covid-19.

Taula 6.6: Resultats utilitzant dos classificadors

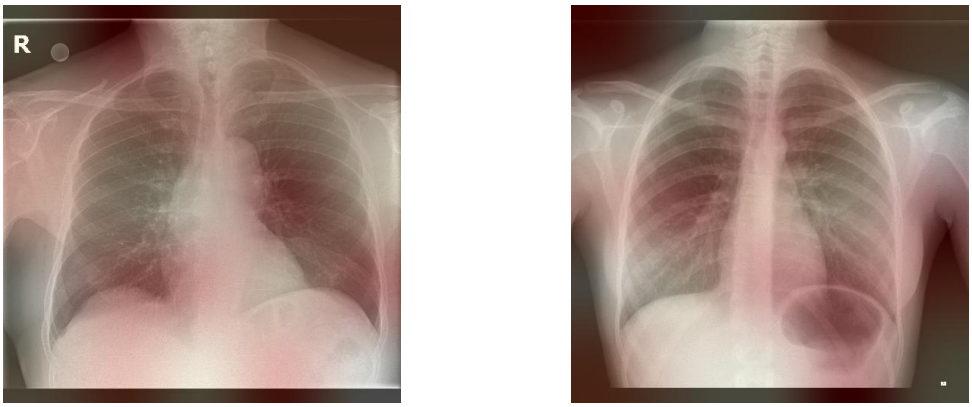
1r model	2n model	Acc.	Precision			Recall			F1-score		
			C	C-19	P	C	C-19	P	C	C-19	P
4A	4A	76,2 %	0,77	0,95	0,63	0,85	0,70	0,64	0,81	0,81	0,64
4B	4B	**78,6 %**	0,80	0,96	0,64	0,83	0,79	0,69	0,82	0,86	0,66
4C	4C	75,2 %	0,73	0,93	0,66	0,88	0,67	0,56	0,80	0,78	0,61
4D	4D	73,9 %	0,76	0,95	0,57	0,81	0,69	0,64	0,79	0,80	0,60
5A	5A	76,2 %	0,76	0,92	0,62	0,86	0,81	0,52	0,81	0,86	0,57
5C	5C	**79,1 %**	0,76	0,97	0,72	0,91	0,77	0,57	0,83	0,86	0,64
7A	7A	61,8 %	0,73	0,00	0,43	0,89	0,00	0,63	0,80	0,00	0,51
4A	5C	77,3 %	0,77	0,96	0,66	0,85	0,74	0,65	0,81	0,84	0,65
4B	5C	**79,2 %**	0,80	0,97	0,65	0,83	0,81	0,70	0,82	0,88	0,67
7A	5C	76,1 %	0,73	0,96	0,71	0,89	0,65	0,61	0,80	0,77	0,66
7A	5A	75,6 %	0,73	0,92	0,71	0,89	0,66	0,58	0,80	0,77	0,64
4B	5A	**78,3 %**	0,80	0,92	0,64	0,83	0,82	0,66	0,82	0,86	0,65
4A	4B	76,2 %	0,77	0,95	0,63	0,85	0,70	0,64	0,81	0,81	0,64
7A	4B	75,9 %	0,73	0,95	0,71	0,89	0,65	0,60	0,80	0,77	0,65
4A	5A	76,4 %	0,77	0,91	0,65	0,85	0,75	0,61	0,81	0,82	0,63
5A	4A	75,7 %	0,76	0,96	0,60	0,86	0,75	0,56	0,81	0,84	0,58
5A	5C	77,2 %	0,76	0,97	0,64	0,86	0,81	0,56	0,81	0,88	0,59
5C	5A	**78,3 %**	0,76	0,92	0,71	0,91	0,78	0,53	0,83	0,84	0,61
5C	4A	**78,0 %**	0,76	0,96	0,69	0,91	0,73	0,56	0,83	0,83	0,62

6.3 Mapes d'activació

En aquesta secció s'analitza el resultat d'aplicar Grad-CAM a imatges de test de cada classe utilitzant els millors models dels classificadors binaris. Tots aquests

mapes d'activació tenen una mida equivalent a l'eixida de l'última capa convolucional de cadascun dels models: 5×5. Per aquest motiu, les zones ressaltades no poden trobar-se molt localitzades sobre la imatge, sinó que simplement orienten a l'observador sobre quina zona ha sigut la més decisiva a l'hora de classificar.

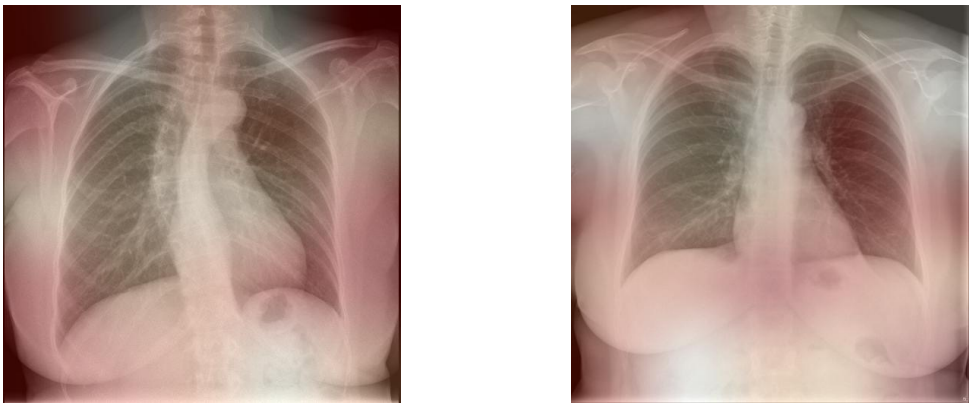
A les imatges de la Figura 6.4 es mostren dos exemples de mapes d'activació superposats a les seues respectives imatges, un per a una imatge de la classe Control i altre per a la classe Pneumonia, que representen aquelles zones que han sigut decisives per al model 7A a l'hora de classificar-les. Tot i que algunes parts no corresponents al pulmó es mostren ressaltades, observem que les parts ressaltades amb més intensitat es troben dintre dels pulmons, sobretot en la imatge de la classe Pneumonia.



(a) Control **(b)** Pneumonia

Figura 6.4: Exemple dels mapes d'activació del model 7A en Control vs. Pneumonia

A les imatges de la Figura 6.5 es mostren els mapes d'activació per a dos imatges classificades pel model 4C en els grups Control i Covid-19. S'observa que les activacions a la imatge de Covid-19 són més intenses a la part dreta de la imatge, dins del pulmó, i també a la part inferior d'ambdós pulmons. Quant a la imatge de Control, les activacions apareixen prou repartides, no centrant l'atenció en cap punt concret.



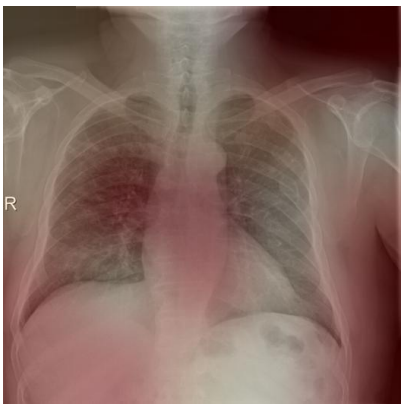
(a) Control **(b)** Covid-19

Figura 6.5: Exemple dels mapes d'activació del model 4C en Control vs. Covid-19

Dels tres parells de mapes d'activació, els de la Figura 6.6 són, potser, els més interessants. Els resultats obtinguts en la classificació entre Pneumonia i Covid-

19 eren els més alts. A la subfigura 6.6b, la qual correspon a una imatge Covid-19, s'observa com les activacions més intenses es troben a la dreta de la imatge, però no dintre del pulmó com caldria esperar. Aquest fet podria indicar que part de les mostres de la classe Covid-19 presenten unes característiques en aquesta zona que no apareixen a les mostres de Pneumonia utilitzades en l'entrenament, i que res tenen a veure amb la malaltia en sí.

(a) Pneumonia



(b) Covid-19

Figura 6.6: Exemple dels mapes d'activació del model 5C en Pneumonia vs. Covid-19

CAPÍTOL 7
Conclusions i treball futur

En aquest projecte s'ha assolit l'objectiu principal que consistia en construir models de xarxes neuronals capaços classificar radiografies segons la presència o no de pneumonia comuna o pneumonia provocada per coronavirus, tot i que l'anàlisi dels resultats, tant de la classificació com de la representació dels mapes d'activació, suggereix que existeixen certes característiques no biomètriques pròpies de les imatges etiquetades com a Covid-19 que alguns dels models arriben a aprendre.

En la classificació entre imatges control i imatges de pneumonia comuna s'arriba a una *accuracy* propera al 80 % i el *f1-score* de les dues classes s'acosta al 0,8. El rendiment de les xarxes augmenta en la classificació entre imatges control i imatges Covid-19, aconseguint fins un 90 % d'*accuracy* i valors al voltant del 0,9 per a l'*f1-score*. Quant a la classificació entre pneumonia comuna i pneumonia per coronavirus, l'*accuracy* supera en alguns casos el 90 %, i el *recall* per a la classe Covid-19 s'acosta a l'1. La millora dels resultats en aquest últim cas respecte dels dos primers fa sospitar que existeixen particularitats pròpies a les imatges de cada classe que faciliten la feina dels models ja que, en un principi, deuria ser més difícil classificar entre dos varietats de pneumonia que no distingir entre la presència o no d'aquesta.

Quant a la classificació en tres classes, s'ha demostrat que dividir el problema en dos classificacions binàries millora els resultats generals. Tot i que un dels models aconsegueix un 80 % d'*accuracy* en la classificació en tres classes, la resta no arriben al 75 %. No obstant, quan es combinen dos classificadors, augmenta l'*accuracy* dels models implicats.

Els mapes d'activació han ajudat a confirmar que els models entrenats se centren en les zones correctes, és a dir, els pulmons, per a classificar entre imatges control i imatges amb algun tipus de pneumonia. Però, en la classificació entre els dos tipus de pneumonia, han ressaltat possibles particularitats entre les imatges de les dues classes que no tenen a veure amb la malaltia en sí. No obstant, per a la distinció entre els dos tipus de pneumonia, continua indicant-se l'interior dels pulmons com una de les zones més decisives en la classificació.

Com a treball futur, es planteja la millora dels mapes d'activació perquè puguen ser més concrets. Per a açò, caldria redissenyar les xarxes ja utilitzades, ja que Grad-CAM genera aquests mapes segons la dimensió de l'eixida de l'última capa convolucional. També seria d'utilitat experimentar amb tècniques d'extrac-

ció de parènquima pulmonar. Finalment, l'objectiu últim seria adaptar les nostres solucions a escenaris reals, seguint les indicacions del personal sanitari especialitzat, per a que els puguen ser de la major utilitat possible.

Bibliografia

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[2] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[3] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017.

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[5] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057, 2015.

[6] Linda Wang and Alexander Wong. COVID-Net: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest X-Ray Images. *arXiv preprint arXiv:2003.09871*, 2020.

[7] Joseph Paul Cohen, Paul Morrison, and Lan Dao. COVID-19 image data collection. *arXiv 2003.11597*, 2020.

[8] Radiological Society of North America. RSNA penumonia detection challenge. 2019.

[9] Zhong Qiu Lin, Mohammad Javad Shafiee, Stanislav Bochkarev, Michael St. Jules, Xiao Yu Wang, and Alexander Wong. Do Explanations Reflect Decisions? A Machine-centric Strategy to Quantify the Performance of Explainability Algorithms. *arXiv*, pages arXiv–1910, 2019.

[10] Ali Narin, Ceren Kaya, and Ziynet Pamuk. Automatic detection of coronavirus disease (Covid-19) using x-ray images and deep convolutional neural networks. *arXiv preprint arXiv:2003.10849*, 2020.

[11] Chest X-Ray Images (Pneumonia). https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia.

[12] B. Ghoshal and A. Tucker. Estimating uncertainty and interpretability in deep learning for coronavirus (COVID-19) detection. arXiv 2020. *arXiv preprint arXiv:2003.10769*.

[13] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, pages 9505–9515, 2018.

[14] Jianpeng Zhang, Yutong Xie, Zhibin Liao, Guansong Pang, Johan Verjans, Wenxin Li, Zongji Sun, Jian He, Yi Li, Chunhua Shen, et al. Viral pneumonia screening on chest X-ray images using confidence-aware anomaly detection. *arXiv: 2003.12338*, 2020.

[15] Juri Yanase and Evangelos Triantaphyllou. A systematic survey of computer-aided diagnosis in medicine: Past and present developments. *Expert Systems with Applications*, 138:112821, 2019.

[16] Gordon D. Schiff, Omar Hasan, Seijeoung Kim, Richard Abrams, Karen Cosby, Bruce L. Lambert, Arthur S. Elstein, Scott Hasler, Martin L. Kabongo, Nela Krosnjar, et al. Diagnostic error in medicine: analysis of 583 physician-reported errors. *Archives of Internal Medicine*, 169(20):1881–1887, 2009.

[17] Nicholas Petrick, Berkman Sahiner, Samuel G. Armato III, Alberto Bert, Loredana Correale, Silvia Delsanto, Matthew T. Freedman, David Fryd, David Gur, Lubomir Hadjiiski, et al. Evaluation of computer-aided detection and diagnosis systems a. *Medical physics*, 40(8):087001, 2013.

[18] Mark A. Helvie, Lubomir Hadjiiski, Erini Makariou, Heang-Ping Chan, Nicholas Petrick, Berkman Sahiner, Shih-Chung B. Lo, Matthew Freedman, Dorit Adler, Janet Bailey, et al. Sensitivity of noncommercial computer-aided detection system for mammographic breast cancer detection: pilot clinical trial. *Radiology*, 231(1):208–214, 2004.

[19] Constance D. Lehman, Robert D. Wellman, Diana S. M. Buist, Karla Kerlikowske, Anna N. A. Tosteson, and Diana L. Miglioretti. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA Internal Medicine*, 175(11):1828–1837, 2015.

[20] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017.

[21] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016.

[22] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. http://tensorflow.org/, 2015. Software available from tensorflow.org.

[23] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pages 265–283, 2016.

[24] François Chollet et al. Keras. https://keras.io, 2015.

APÈNDIX A
Models

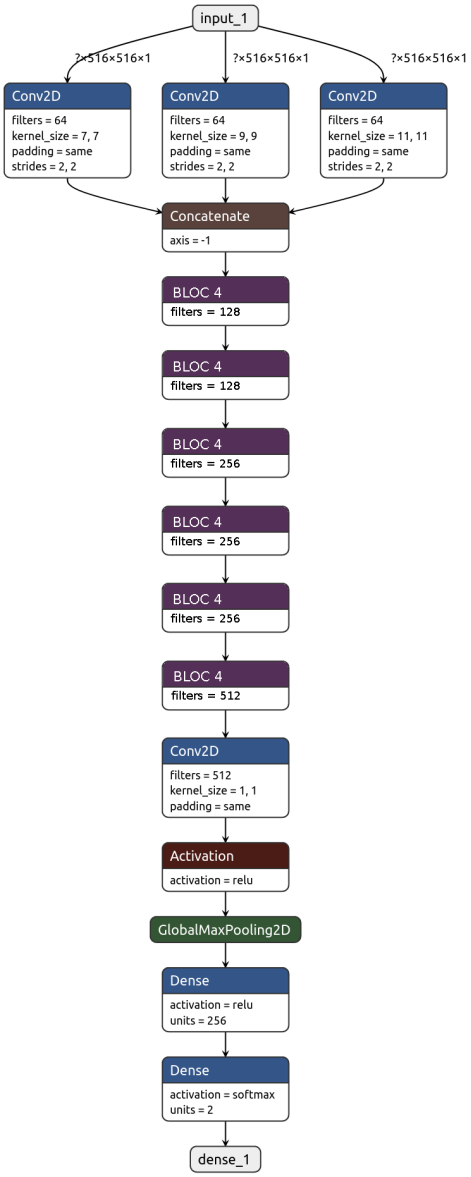


Figura A.1: Model 4A

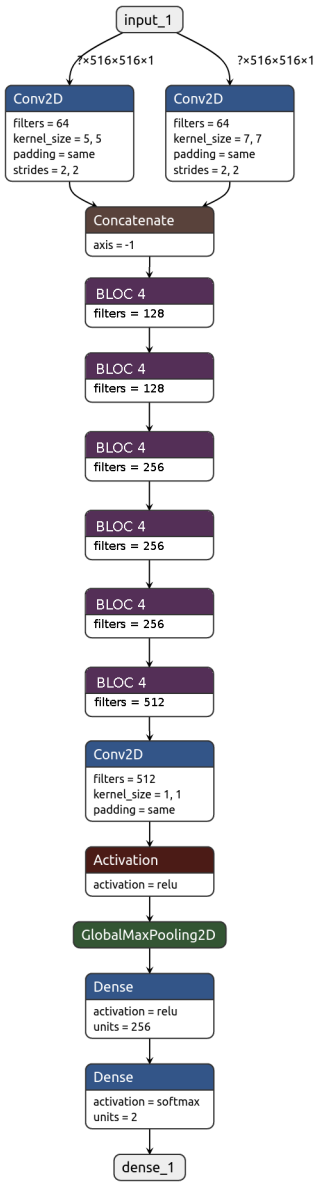


Figura A.2: Model 4B

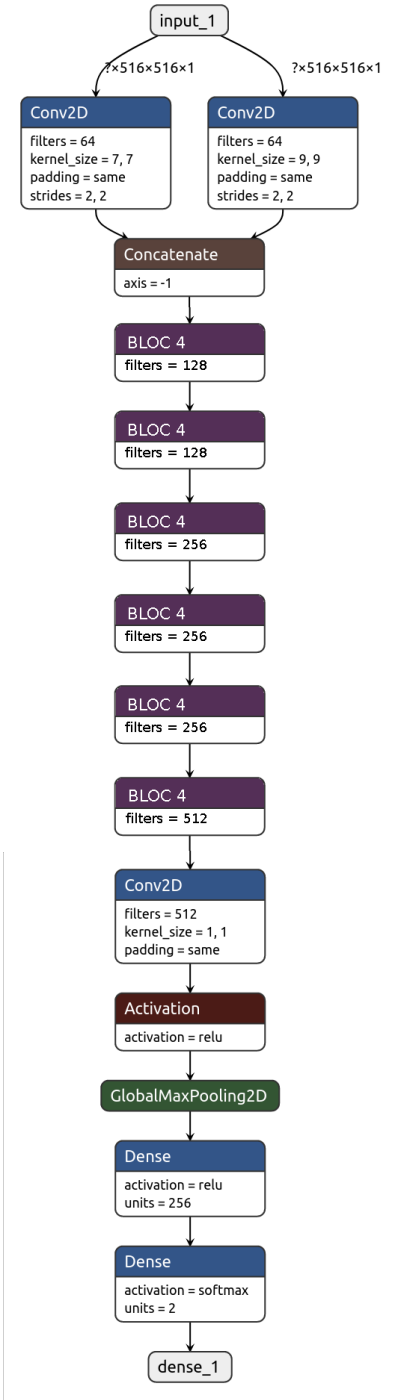


Figura A.3: Model 4C

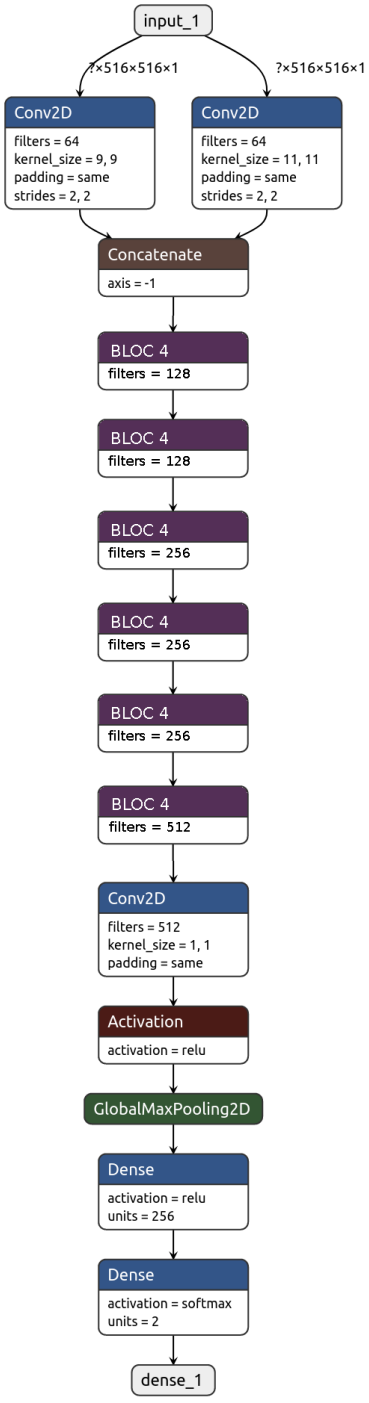


Figura A.4: Model 4D

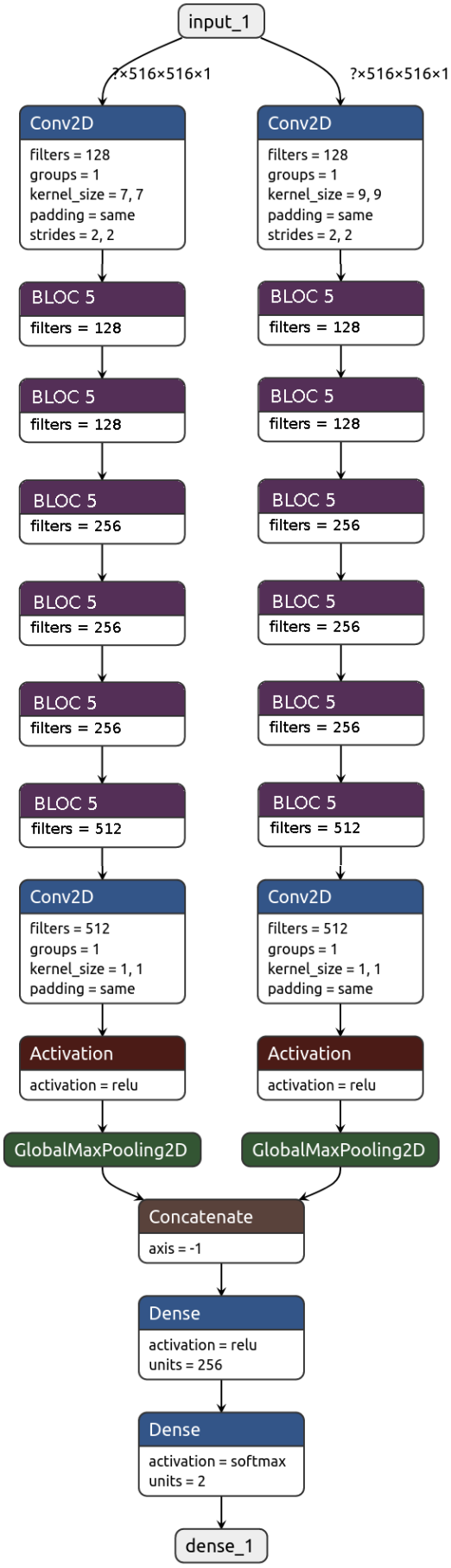


Figura A.5: Model 5A

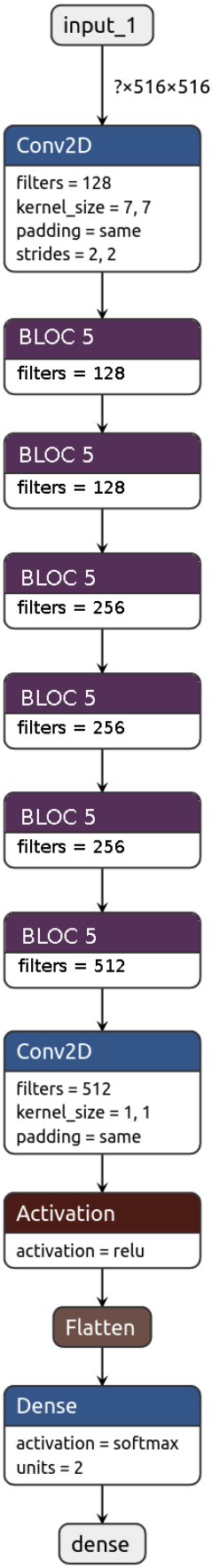


Figura A.6: Model 5B

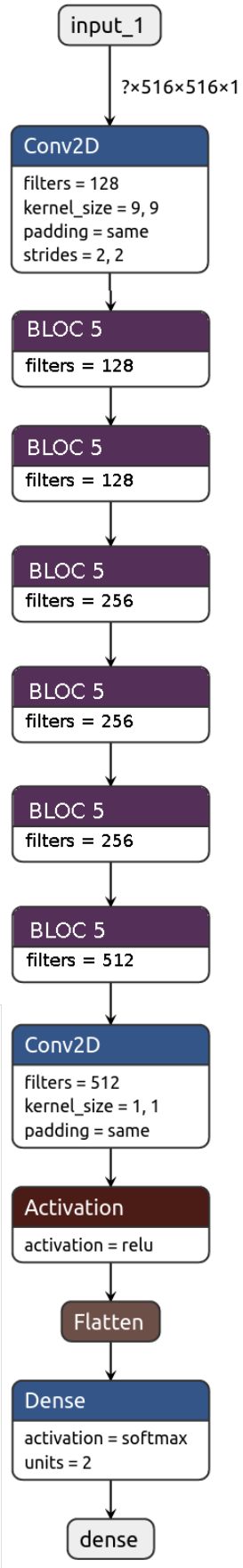


Figura A.7: Model 5C

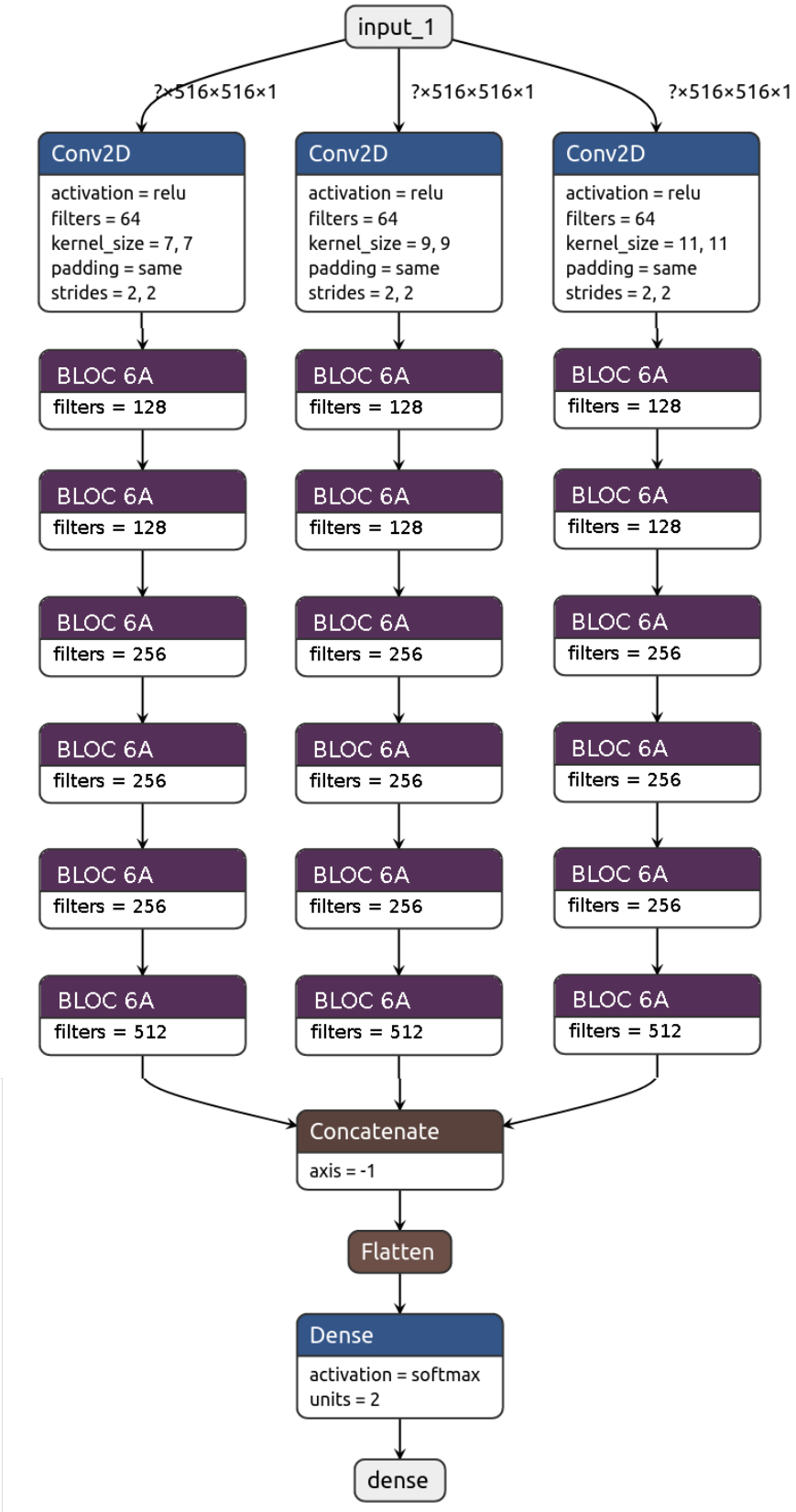


Figura A.8: Model 6A

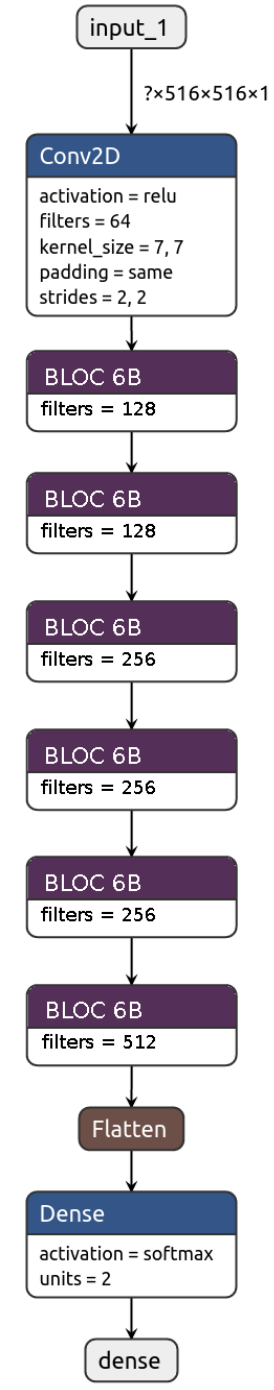


Figura A.9: Model 6B

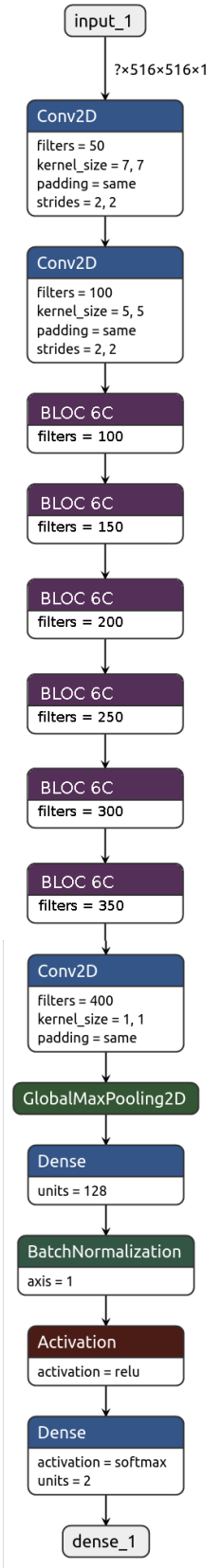


Figura A.10: Model 6C

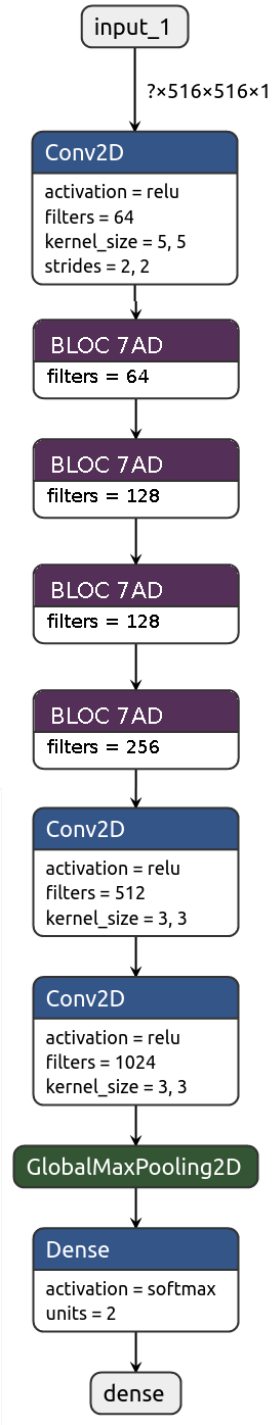


Figura A.11: Model 7A

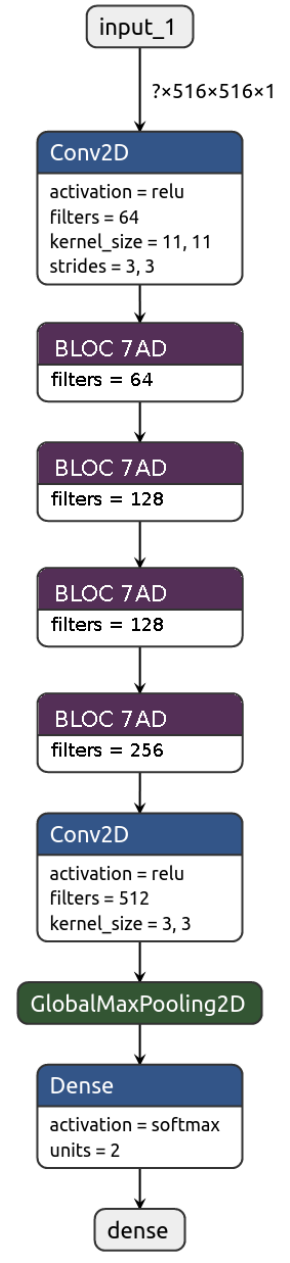


Figura A.12: Model 7D

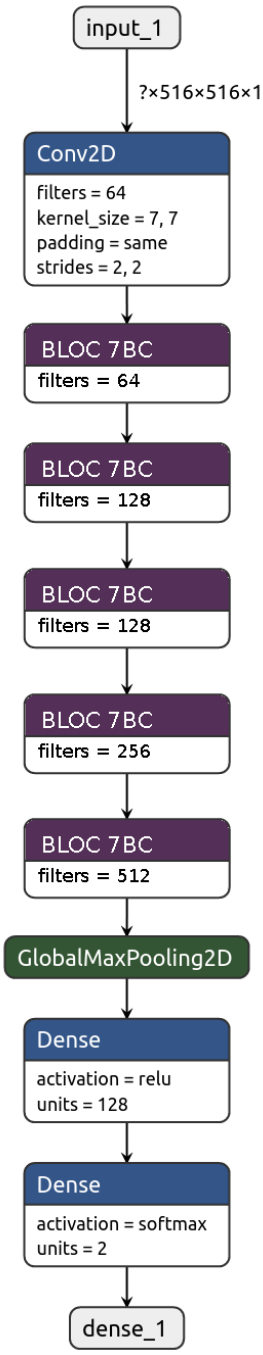


Figura A.13: Model 7B

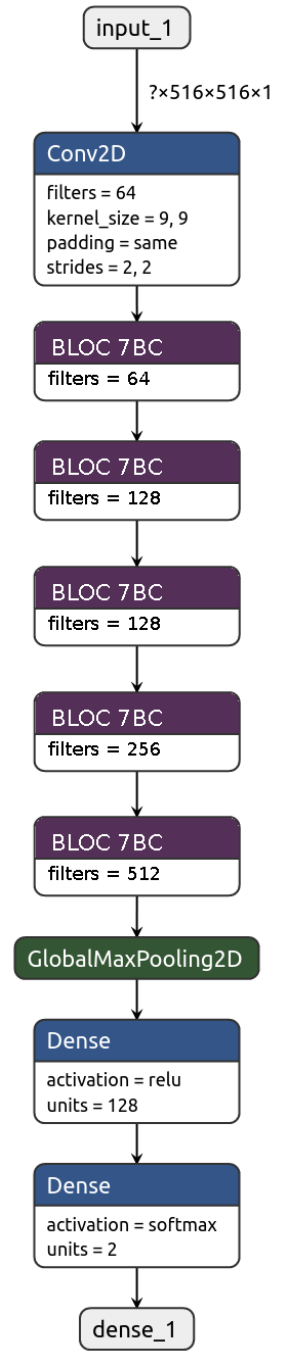


Figura A.14: Model 7C