Potenciális COVID-19 fertőzés automatikus felismerésé hagyományos véranalízis alapján

Automatic detection of potential COVID-19 infection based on conventional blood analysis

Zoltán CZAKO[1], Gheorghe SEBESTYEN[2], Anca HANGAN[3]

Számítástechnikai Tanszék, Kolozsvári Műszaki Egyetem, Koloszvár, Románia
[1]zoltan.czako@cs.utcluj.ro, [2]gheorghe.sebestyen@cs.utcluj.ro, [3]anca.hangan@cs.utcluj.ro

Abstract

To control the spread of the COVID-19 it is very important to identify those who have been already infected by this new type of virus. The rRT-PCR (reverse transcription polymerase chain reaction) testing is the golden standard for COVID-19 detection, but it is time consuming, laborious manual process and it is very short in supply. To reduce the number of tests, in this article we will present a possible solution for COVID-19 preliminary patient filtering based on regular blood tests, using artificial intelligence (AI) models. The most appropriate AI model will be selected using our auto-adaptive AI platform, AutomaticAI. The hyperparameters of the selected algorithm will also be adjusted automatically by this platform to match the context of the problem.

Keywords: COVID-19, Regular Blood Test, AutomaticAI, Particle Swarm Optimization, Simulated Annealing

Kivonat

A COVID-19 terjedésének megfékezése érdekében nagyon fontos azonosítani azokat a személyeket, akiket már megfertőzött ezen új típusú vírus. Az rRT-PCR (reverse transcription polymerase chain reaction) teszt a COVID-19 detektálásának leghatékonyabb eszköze, ám időigényes, fárasztó kézi folyamat, és nagyon szűk a készlet belőle. A tesztek számának csökkentése érdekében, ebben a cikkben a COVID-19 előzetes betegszűrésének lehetséges megoldását mutatjuk be hagyományos vérvizsgálatok alapján, mesterséges intelligencia (AI) modellek felhasználásával. A leghatékonyabb AI-modellt automatikusan alkalmazkodó AI-platformunk, az AutomaticAI segítségével választjuk ki. A kiválasztott algoritmus hiperparamétereit platformunk képes automatikusan beállítani, ezáltal megfelelve a probléma kontextusának.

Kulcsszavak: COVID-19, vérteszt, AutomaticAI, Részecske Raj Optimalizálás, Szimulált Hűtés

1. Bevezető

A 2019-es koronavírus-betegséget (COVID-19) 2019 decemberében azonosították Wuhanban, Kínában [3], ami a 2019–2020-as járványt eredményezte. Ez az új SARS-CoV-2 vírus nagyon agresszív, mivel sokkal könnyebben terjed az emberek között, mint a legközelebbi rokona, a SARS. A számítógépes modellezési és szimulációs technikák arra utalnak, hogy minden új COVID-19 eset átlagosan 2,67 más személyt fertőz meg [4]. Az új vírus elleni küzdelem leghatékonyabb módja a társadalmi távolságtartás, a gyors és korai észlelés és izolálás. A korai felismerés nagyon nehéz lehet, mivel az inkubációs periódus (a fertőzés és a megjelenő tünetek között) 2 és 14 nap között lehet (az Egészségügyi Világszervezet szerint 5 nap a leggyakoribb). A fertőzés sok esetben tünetmentes (fertőzött, de nincs jele vagy tünete), vagy egyes betegeknél influenza-szerű tünetek, láz, köhögés vagy légszomj alakul ki. További súlyosbodó tünetek a mellkasi fájdalom vagy nyomás. A SARS-CoV-2-vel fertőzött egyének

esetenként más tüneteket, például tüsszentést, orrfolyást, hányást, hasmenést vagy torokfájást mutathatnak. Mivel a tünetek hasonlóak lehetnek más légzőszervi fertőzésekhez vagy influenzához, a korai észlelés nehéz; nehéz különbséget tenni a COVID-19 és az influenza között, csak a tünetek alapján. A Koronavírus kimutatásának másik módja a polimeráz láncreakció (PCR) [1] vagy valós idejű reverz transzkripciós polimeráz láncreakció (rRT-PCR) [2] eljárás használata. Az ilyen típusú teszteknél az a probléma, hogy lassúak és speciális berendezéseket és vegyszereket igényelnek, amelyek nehezen beszerezhetőek.

Ilyen körülmények között az orvosoknak gondosan meg kell választaniuk, hogy kit kell tesztelni és kit nem. Az egyik megoldás az embercsoportok létrehozása és csoportonként egy teszt elvégzése (keverjük össze a mintákat, majd csak egy PCR-tesztet végezzünk). Ha a teszt negatív, akkor az egész csoportot csak egy teszttel sikerült kipróbálni. Ha a teszt pozitív, válasszuk a csoportot két részre, és kövessük a bináris keresési szabályokat. A tesztelésre kerülő emberek kiválasztásának hatékonyabb módja más, olcsóbb és gyakoribb tesztek használata a SARS-CoV-2 fertőzés kockázatának felmérésére. Ebben a cikkben egy egyszerű vérvizsgálaton alapuló COVID-19 előzetes szűrési algoritmust mutatunk be, ami hatékony lehet a PCR tesztre szoruló emberek kiválasztása érdekében.

Ez az algoritmus segíthet az orvosoknak kiválasztani a tesztre szoruló embereket, viszont nem helyettesíti a PCR tesztet, csak kiegészíti azt. Algoritmusunk előzetes tesztként használható, mielőtt a COVID-19 detektálási teszteket alkalmaznánk. Ha algoritmusunk eredménye azt mutatja, hogy a fertőzés esélye nagyon alacsony vagy nulla százalék, akkor a beteg egészségesnek tekinthető, és további vizsgálatokra nincs szükség. Máskülönben további vizsgálatokra van szükség (vagyis COVID-19 vizsgálatára van szükség). Tehát ez az algoritmus automatikus előszűrésnek tekinthető, a potenciálisan vírussal fertőzött személyeket (ideértve a SARS-CoV-2-t is) határolja el az egészséges személyektől.

A cikk többi része a következőképpen van felépítve: a 2. részben néhány érdekes megoldást mutatunk be a COVID-19 észlelésére és osztályozására, valamint összehasonlítjuk azokat más hasonló munkákkal, a 3. részben leírjuk a COVID-19 előzetes szűrésére és az eredmények magyarázatára használt algoritmust, a 4. részben a kísérleti eredményeket mutatjuk be, az 5. szakaszban pedig az összefoglalót találhatjuk.

2. Kapcsolódó munkák

A jelenlegi pandémiás helyzet sok kutatást generált a COVID-19 osztályozása és kimutatása területén. Számos kutató megkísérel hagyományos tradicionális értékelési technikákat (analíziseket), például röntgen, CT vagy vérvizsgálatot használni a COVID-19 fertőzés azonosítására. Az egyik legnépszerűbb és legvitatottabb probléma a SARS-CoV-2 fertőzések kimutatása CT vagy X-Ray tüdőképek alapján.

A [5] cikkben két különféle megközelítést láthatunk a COVID-19 osztályozására CT képek alapján. Először, a szerzők megpróbálták a képeket osztályozni tulajdonságkivonás (feature extraction) nélkül, a képeket vektorokká alakították át és az osztályozást SVM algoritmussal végezték el [12]. Mivel az eredmények nem voltak túl ígéretesek, egy másik megközelítést próbáltak ki, amelyben az SVM alkalmazása előtt öt különféle tulajdonságkivonási algoritmust használtak, nevezetesen a “Grey Level Co-occurrence Matrix”, “Local Directional Pattern” , “Grey Level Run Length Matrix” , “Grey Level Size Zone Matrix” (GLSZM) [13] és diszkrét hullámtranszformációt (Discrete Wavelet Transform). Az egyes tulajdonságkivonási algoritmusokkal végzett kísérletek elvégzése után, a legjobb eredményt a GLSZM alkalmazásával kapták, amely 13 jellemző vektort (feature vector) generált. A GLSZM eredményeit SVM-mel használták a képek osztályozása érdekében. Így sikerült 96,46 + -3,7% -os f1-értéket kapni.

A [6] cikkben nagyon hasonló megközelítést láthatunk, amelyet az X-Ray képekre alkalmaztak. A különbség az, hogy mielőtt az [5] cikkben említett öt tulajdonságkivonási algoritmust felhasználták, a SMOTE algoritmust használták a túlmintavételhez (over-sampling) (mivel a COVID-19-re csak néhány röntgenkép van, és sok példa a tüdőgyulladás és az egészséges betegek adatait tartalmazza), és a tulajdonságkivonási algoritmusok alkalmazása után PCA-t alkalmaztak dimenziócsökkentésre, így a kapott vektorok dimenziója 78 jellemzőről 20 jellemzőre csökkent. A kapott tulajdonságvektorokra SVM-et alkalmaztak, ily módon sikerült 84,5 százalékos f1-értéket elérni, ami az eddig elért egyik legjobb eredmény a konvolúciós neurális hálózatok (CNN) használata nélkül.

Más cikkekben, mint például a [7] cikkben, újonnan létrehozott konvolúciós neurális hálókat találhatunk egyedi rétegekkel, mint például a vetítés (projection), a kiterjesztés (expansion) és a mélység szerinti ábrázolás (depth-wise representation). Az újonnan létrehozott COVID-Net architektúrával a szerzők 80% -os pozitív prediktív értéket és majdnem 100% érzékenységet (sensitivity) kaptak. A [8] cikkben három különféle CNN architektúrát teszteltek átviteli tanulással (transfer learning). Ebben az esetben a legjobb eredményt a ResNet50 alkalmazásával kapták, de a szerzők az InceptionV3 [10] és az Inception-ResNetV2 architektúrát is tesztelték, amelyek pontossága 87, illetve 97 százalék volt.

Az összes fent említett cikkben megoldásokat találhatunk a COVID-19 osztályozására különféle típusú képek (CT vagy röntgen) alapján. A COVID-19 betegség első, tünetmentes stádiumában, amikor a vírus fertőzési valószínűsége nagy, a vírus tüdőkre gyakorolt hatása meglehetősen csekély. Ezért a tüdőkép elemzése túl későn történik, tehát nem a legmegfelelőbb eszköz a vírus korai észlelésére és megelőzésére. Ezért ebben a cikkben a fertőzés megelőzésének problémáját más szögből fogjuk vizsgálni; hétköznapi vérvizsgálatok eredményeit fogjuk felhasználni a fertőzés valószínűségének becslésére. Az algoritmus felhasználható előszűrőként annak kiválasztására, akit tesztelni kell, ezáltal csökkentve a szükséges specifikus COVID-19 tesztek számát, és segítve az orvosokat abban, hogy automatikusan különválaszthassák a magas fertőzésveszélyes betegeket.

3. A COVID-19 előzetes szűrése az AutomaticAI platform felhasználásával

3.1. Automatikus előzetes szűrés

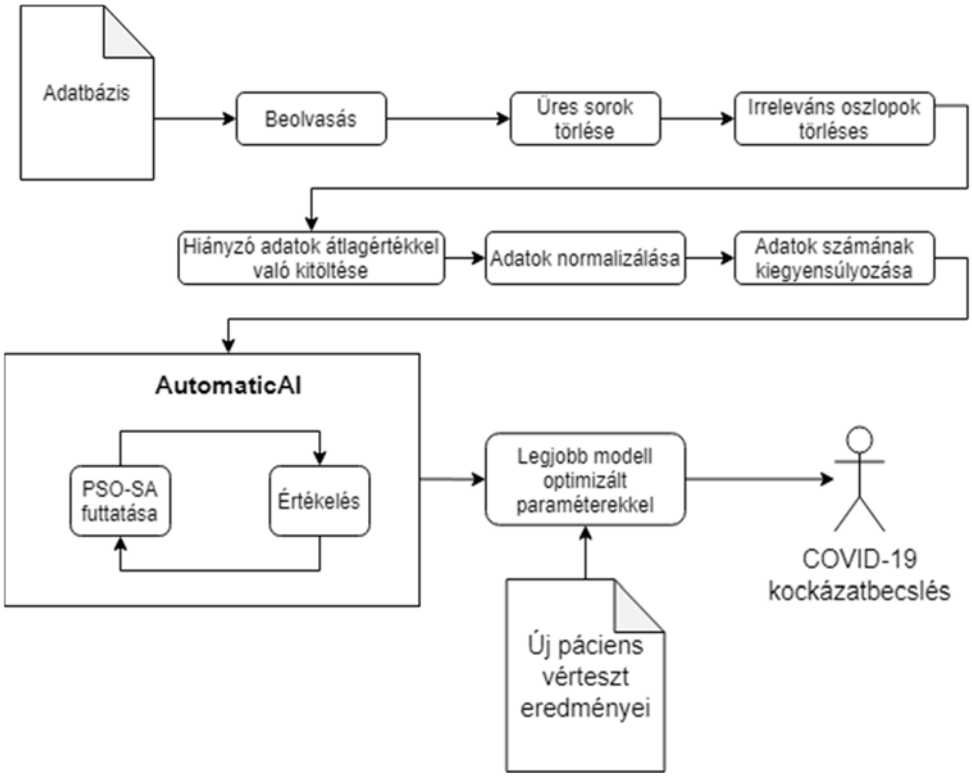
A munka fő célja egy mesterséges intelligencia algoritmus létrehozása, amely segíthet az orvosoknak abban, hogy automatikusan kiszűrjék az egészséges betegeket, és kiválasszák azokat a betegeket, akik esetében további vizsgálatok vagy COVID-19 tesztek szükségesek. Elképzelésünk az, hogy hétköznapi vérvizsgálati eredményeket használhatunk az egészséges betegek kiszűrésére és a vírusfertőzés valószínűségének becslésére. Ha van fertőzés, akkor további vizsgálatot kell végezni annak ellenőrzése érdekében, hogy ezt a SARS-CoV-2 és nem más típusú vírus okozza. Első célunk tehát az egészséges és vírusfertőzésre hajlamos betegek osztályozása, ideértve a COVID-19-et is. A kutatás ezen szakaszában módszerünk nem pótolja a PCR COVID-19 tesztek szükségességét, inkább előzetes szűrőként, sokkal olcsóbb és gyorsabb osztályozási módszerként értendő.

A bevitt adatkészlet alapján a legjobb modell és a legjobb hiperparaméter-konfiguráció megtalálásához az 1.-es ábrán bemutatott megoldást használtuk. Amint azt az 1.-es ábra mutatja, az első három lépésben a hiányos, hiányzó vagy irreleváns adatok problémáját a következőképpen kezeltük:

1. Töröltük azokat a sorokat, ahol a vérvizsgálatok teljesen hiányoztak
2. Eltávolítottuk azokat az oszlopokat, amelyek nem relevánsak a besorolási probléma szempontjából (tehát minden oszlopot eltávolítottunk, amelyek olyan adatokat tartalmaztak, amelyeket nem lehet csakis hétköznapi vérvizsgálattal mérni, hanem más vizsgálatok is szükségesek)
3. Azon oszlopok esetén, amelyekben csak néhány hiányzó adat volt, kitöltöttük azokat az adott oszlop átlagértékének felhasználásával.

A következő lépésekben átalakítottuk (normalizáltuk) az adatokat, hogy mindegyik tulajdonság (oszlop) azonos nagyságrendű legyen. Ezután a kiegyensúlyozatlan adatok problémáját a 0.4-es tényezővel történő túlmintavétel (oversampling) segítségével kezeltük.

A megtisztított és átalakított adatokat az AutomaticAI platform számára [14] bemeneti adatként használtuk, itt futtattuk a PSO-SA algoritmust, hogy kiválasszuk a legjobb osztályozási algoritmust és a kiválasztott algoritmusnak megfelelő legjobban teljesítő paraméter beállítást. Ezen platform számos olyan AI osztályozási modell tesztelését végzi, amelyek különböző paramétereket tartalmaznak, amíg meg nem találja a legjobban teljesítő modelleket és legjobb hiperparaméter beállításokat.



1. ábra. *A megoldáshoz vezető lépések*

3.2. Felhasznált adatkészlet

A kísérleteinkben használt adatkészlet [11] anonim egészségügyi nyilvántartást tartalmaz a potenciális COVID-19-fertőzött betegekről, akiket a brazil Sao Pauloban, valamint az Israelita Albert Einstein kórházban láttak el. Az összes adatot névtelenítették a legjobb nemzetközi gyakorlatok és ajánlások alapján. Az összes klinikai adatot úgy standardizálták, hogy nulla átlaga és egységnyi szórása legyen. A dátumokat kihagyták, és a beteg nemére vonatkozó információkat kódolták. Ez az adatkészlet több mint 5000 beteg adatait tartalmazza, és minden páciens több mint 100 tulajdonsággal rendelkezik. A jellemzők között hétköznapi vérvizsgálat eredményeit is tartalmazza, olyan jellemzőkkel, mint hematokrit, hemoglobin, vérlemezkék, átlagos vérlemezke-térfogat, vörösvértestek, limfociták, MCHC (átlagos korpuszkuláris hemoglobin-koncentráció), leukociták, basophilok, MCH (átlagos korpuszkuláris hemoglobin), stb. A hétköznapi vérvizsgálati értékek mellett a vírusok különböző tesztjeinek eredményeit is tartalmaz, például légzőszervi szintetikus vírus, A-influenza, B-influenza, orrszarvú / enterovírus, 1., 2., 3. influenza, Adenovírus stb. Továbbá ezen adatkészlet tartalmaz egy bináris oszlopot is, amely a rRT-PCR teszt eredményeit tartalmazza, mely esetben a negatív érték egészséges személyre, míg a pozitív értek Koronavírussal fertőzött betegre utal.

3.3. Az AI modell kiválasztásához és optimalizálásához használt algoritmus

A legjobb mesterséges intelligencia osztályozási modell megkeresése és a hiperparaméterek optimizálása céljából egy hibrid PSO-SA (Particle Swarm Optimization - Simulated Annealing) algoritmust használtunk az AutomaticAI platformon. Ez egy általános célú algoritmus, amelyet bármilyen osztályozási vagy predikciós problémában fel lehet használni, képes arra, hogy automatikusan beállítsa magát, és kiválasszon egy megfelelő algoritmust az aktuális kontextus és a bemeneti adatok alapján. Ha ezt az algoritmust kombináljuk az előfeldolgozási és a jellemzőkivonási (feature extraction) lépésekkel, szinte bármilyen osztályozási (vagy regressziós) problémát meg lehet oldani meglehetősen magas pontszámokkal (pontosság, recall stb.). Ennek az algoritmusnak az az előnye, hogy nem kell manuálisan keresnünk az AI modelleket és azokat egyenként értékelnünk, nem kell manuálisan keresnünk a hiperparaméterek értékeit, amely (manuálisan) nagyon időigényes munka és néha lehetetlen, figyelembe

véve, hogy néha a hiperparaméter-kombinációk száma nagyon magas (végtelen is lehet multidimenziós térben). A platformunkba beépített PSO-SA algoritmus esetén az algoritmustípusonkénti részecskék száma manuálisan is beállítható, és szimulált hűtéshez hasonló heurisztikát használ a lokális minimális vagy maximális értékek elkerülésére (attól függően, hogy hibát minimalizálunk vagy pontosságot maximalizálunk).

Ebben az algoritmusban minden részecske egy algoritmustípust képvisel, és több raj/részecskecsoport létezik (algoritmustípusonként egy raj). Mindegyik csoportnak van egy vezetője, azon részecske, amely a legmagasabb értékelési pontszámmal rendelkezik, és csak egy globális vezető van, azon részecske, amely az általános (nem csak csoportonként) legjobb mutatóval rendelkezik. Az egyes részecskék mozgását befolyásolja a helyi vezető, a globális vezető és a személyes legjobb pozíció. A következő pozíció elfogadási kritériumába egy Szimulált Hűtéshez hasonló kritériumot is alkalmaztunk: először a részecskék nagyobb mozdulatokat hajtanak végre (egyesek rossz irányba), megpróbálva elkerülni a helyi minimumokat vagy maximumokat; az iterációk számának növekedésével a lépés nagysága fokozatosan csökken, az optimális megoldás pontosabb megközelítése érdekében.

Minden korszakban eltávolítjuk a legkisebb pontszámú részecskéket, és ugyanannyi új részecske kerül azon csoportba, ahol a globális legjobb részecske megtalálható, mint amennyi részecskét eltávolítottunk. Így oldottuk meg a modellkiválasztás problémáját, mivel néhány korszak után minden részecske megkeresi ugyanazon algoritmustípus hiperparamétereit. Ezzel a módszerrel a hiperparaméterek hangolását implicit módon megoldottuk, hisz a részecske helyzetvektorának minden eleme egy adott algoritmus hiperparamétereit jelképezi.

4. Kísérleti eredmények

A PSO-SA algoritmus egy optimalizáló algoritmus, amely bizonyos heurisztikákat alkalmaz, tehát minden egyes futtatás kissé eltérő eredményeket fog adni. Annak igazolására, hogy az algoritmus minden futtatás után nagyon hasonló eredményeket ad vissza, többször futtattuk, és minden alkalommal 97 + -2,0% f1-értéket kaptunk, ami viszonylag stabil eredmény.

	Pontosság	**F1-érték**	**AUC**	**Algoritmus és paraméterei**
1.	0.942	0.968	0.978	RandomForestClassifier(criterion='gini', min_samples_split=8, n_estimators=155)
2.	0.956	0.956	0.976	ExtraTreesClassifier(criterion='gini', min_samples_split=2, n_estimators=186)
3.	0.987	0.982	0.983	ExtraTreesClassifier(criterion='gini', min_samples_split=2, n_estimators=199)

1. Táblázat. Az algoritmus eredményei

Az első három modell a hiperparaméterekkel és az eredményekkel az 1. táblázatban látható. Mindegyik esetben ugyanazzal a számú részecskével és azonos hardverkonfigurációval futtattuk az algoritmust.

2. ábra. *Összetévesztési mátrix (Confusion Matrix)*

A legjobb modell összetévesztési mátrixát (confusion matrix) a 2.-es ábra szemlélteti.

5. Következtetések

Ebben a cikkben bemutattunk egy lehetséges megoldást a Koronavírussal fertőzött betegek kiszűrésére Mesterséges Intelligencia alkalmazásával. A cikk első részében ismertettük azokat a lépéseket, amelyek szükségesek a bemeneti adatkészlet tisztításához, a hiányzó adatok kezeléséhez és a kiegyensúlyozatlan adatok káros hatásainak enyhítéséhez. A második részben bemutattuk a hibrid PSO-SA algoritmust, amely egy általános megoldás a kontextusnak legmegfelelőbb mesterséges intelligencia modell automatikus kiválasztására és a hiperparaméterek optimalizálására. Ezt az algoritmust arra használtuk, hogy automatikusan megtaláljuk a bemeneti adatoknak legmegfelelőbb AI modellt, és beállítsuk annak hiperparamétereit, ezáltal maximizálva az értékelési mutatókat (például a pontosság vagy az AUC).

Az algoritmus többszöri futtatása után bemutattuk a három legjobb eredményt elért modellt, amelyek közül kiválasztottuk a lehető legjobbat, melyet az egészséges és a potenciálisan fertőzött betegek elkülönítésére használtuk fel. A legmagasabb pontszámot mutató algoritmus az Extra Tree Classifier nevezetű algoritmus volt, amely 98,2%-os f1-értéket ért el, 98,7% -os pontossággal, ami igazán kiemelkedő eredménynek számít.

Hivatkozások

1. Wang et al. Detection of sars-cov-2 in different types of clinical specimens, JAMA, 2020.
2. Udugama B, Kadhiresan P, Kozlowski HN, Malekjahani A, Osborne M, Li VYC, Chen H, Mubareka S, Gubbay JB, Chan WCW. Diagnosing COVID-19: The Disease and Tools for Detection, ACS Nano, 9th of April, 2020
3. Wu Yi-Chia, Chen Ching-Sunga, Chan Yu-Jiuna The outbreak of COVID-19, Journal of the Chinese Medical Association, March 2020, Volume 83, Issue 3, p. 217-220
4. Ying Liu, Albert A Gayle, Annelies Wilder-Smith, Joacim Rocklov, The reproductive number of COVID-19 is higher compared to SARS coronavirus, Journal of Travel Medicine, March 2020, Volume 27, Issue 2
5. Barstugan Mucahid, Ozkaya Umut and Ozturk Saban, Coronavirus (COVID-19) Classification using CT Images by Machine Learning Methods, arXiv, March, 2020
6. Barstugan Mucahid, Ozkaya Umut and Ozturk Saban, Classification of Coronavirus Images using Shrunken Features, arXiv, April, 2020
7. S.U.K. Bukhari, The diagnostic evaluation of Convolutional Neural Network (CNN) for the assessment of chest X-ray of patients infected, medRxiv, 26th of March, 2020
8. 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, April, 2020
9. Ali Narin, Ceren Kaya and Ziynet Pamuk, Automatic Detection of Coronavirus Disease (COVID-19) Using X-ray Images and Deep Convolutional Neural Networks, arXiv, March, 2020
10. Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens and Zbigniew Wojna, Rethinking the Inception Architecture for Computer Vision, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016
11. Kaggle, 2020, Einstein Data4u, accessed 22th of April, 2020, https://www.kaggle.com/einsteindata4u/covid19/version/4
12. Adankon Mathias M. and Cheriet Mohamed, Support Vector Machine, Springer US, 2009
13. Guillaume Thibault, Bernard Fertil, Claire Navarro, Sandrine Pereira, Pierre Cau, Nicolas Levy, Jean Sequeira and Jean-Luc Mari, Texture indexes and gray level size zone matrix. Application to cell nuclei classification, 10th International Conference on Pattern Recognition and Information Processing, 2009
14. Czako Zoltan, AutomaticAI - Platform for Automatic Evaluation, Optimization and Selection of Artificial Intelligence Algorithms, Dissertation, Technical University of Cluj-Napoca, 2019