

GÉPI TANULÁSI MÓDSZEREK ALKALMAZÁSA A VULKANOLÓGIA ÉS MAGMÁS GEOKÉMIA TERÜLETÉN

SZAKDOLGOZAT FÖLDTUDMÁNYI ALAPSZAK GEOLÓGUS SZAKIRÁNY

Készítette: Pánczél Emese

Témavezető: Dr. Haranginé Lukács Réka Zsuzsanna

MTA-ELTE Vulkanológiai Kutatócsoport

Konzulens: Dr. Csabai István

ELTE TTK Komplex Rendszerek Fizikája Tanszék

**Eötvös Loránd Tudományegyetem
Földrajz- és Földtudományi Intézet
Kőzettan-Geokémiai Tanszék**



2021

NYILATKOZAT

Név: Pánczél Emese

ELTE Természettudományi Kar, szak: Földtudományi Alapszak

NEPTUN azonosító: GD61WD

Szakdolgozat címe:

Gépi tanulási módszerek alkalmazása a vulkanológia és magmás geokémia területén

A **szakdolgozat** szerzőjeként fegyelmi felelősségem tudatában kijelentem, hogy a dolgozatom önálló szellemi alkotásom, abban a hivatkozások és idézések standard szabályait következetesen alkalmaztam, mások által írt részeket a megfelelő idézés nélkül nem használtam fel.

Budapest, 2021. 05. 15.



a hallgató aláírása

Tartalomjegyzék

1.	Bevezetés	4
2.	Gépi tanulás	6
2.1.	A gépi tanulási folyamat lépései	7
2.1.1.	Adatbázisépítés.....	7
2.1.2.	Adathalmaz szétoztása.....	8
2.1.3.	Adattisztítás, dimenziócsökkentés, jellemzők kinyerése	8
2.1.4.	Tanítás	10
2.1.5.	Validálás.....	12
2.1.6.	Tesztelés	13
2.2.	A gépi tanulás főbb módszerei	13
3.	Gépi tanulás a vulkanológiában és magmás geokémiában.....	21
3.1.	Aktív vulkánok valós idejű megfigyeléséből származó adatok kiértékelése	22
3.1.1.	Vulkáni földrengések	23
3.1.2.	Egyéb vulkán monitoring idősorok és multiparaméteres modellek	30
3.2.	„Statikus” magmás geokémiai adatok kiértékelése.....	39
4.	A gépi tanulás értékelése magmás geokémiai-vulkanológiai szemszögből	46
5.	Összefoglalás	49
6.	Köszönetnyilvánítás	49
7.	Irodalomjegyzék	50

A dolgozatban gyakran alkalmazott rövidítések és jelentésük jegyzéke

AI	artificial intelligence	mesterséges intelligencia
ANN	artificial neural network	mesterséges neurális háló
CA	cluster analysis	klaszteranalízis
CCDA	combined cluster and discriminant analysis	kombinált klaszter- és diszkriminancia analízis
CNN	convolutional neural network	konvolúciós neurális háló
DL	deep learning	mély tanulás
DNN	deep neural network	mély neurális háló
DT	decision tree	döntési fa
FA	factor analysis	faktoranalízis
HMM	hidden Markov model	rejtett Markov-modell
ICA	independent component analysis	független komponens analízis
IR	infrared	infravörös
KNN	k-nearest neighbours	k-legközelebbi szomszéd osztályozó
LDA	linear discriminant analysis	lineáris diszkriminancia-analízis
LP	long-period earthquake	hosszú periódusú földrengés
LPC	linear predictive coding	lineáris prediktív kódolás
LR	logistic regression	logisztikus regresszió
MAE	mean absolute error	átlagos abszolút hiba
MDS	multidimensional scaling	többdimenziós skálázás
ML	machine learning	gépi tanulás
MLP	multilayer perceptron	többrétegű perceptron
MSE	mean squared error	átlagos négyzetes hiba
NB	naive Bayes classifier	naiv Bayes osztályozó
NNMF	non-negative matrix factorization	nemnegatív mátrixfaktORIZÁCIÓ
PCA	principal component analysis	főkomponens analízis
RF	random forest	véletlen erdő
RMSE	root-mean-square error	átlagos négyzetes hiba négyzetgyöke
RNN	recurrent neural network	visszacsatolt neurális háló
SOM	self-organizing map	önszervező térkép
SVD	singular value decomposition	szinguláris érték felbontás
SVM	support vector machine	tartóvektor-gép
SWIR	short-wave infrared	rövidhullámú infravörös
ULF	ultra low frequency	ultra alacsony frekvencia
VLP	very long period earthquake	nagyon hosszú periódusú földrengés
VSR	volcano-seismic recognition	vulkáni földrengések valós idejű kiértékelése
VT	volcanotectonic earthquake	vulkanotektonikai földrengés

1. Bevezetés

A mesterséges intelligencia (artificial intelligence, AI) korunk rohamosan fejlődő területe (Goodfellow et al. 2016), melyet Andrew Moore, a Carnegie Mellon Egyetem Számítástudományi Tanszékének volt dékánja úgy definiált, mint „az a tudomány és mérnöki technológia, amely lehetővé teszi, hogy a számítógépek úgy viselkedjenek, amiről korábban azt gondoltuk, hogy emberi intelligenciát igényel” (Carniel & Guzmán 2020). A mesterséges intelligencia részét képező gépi tanulás (machine learning, ML) olyan algoritmusok tudománya, amelyek segítségével a számítógépes programok tapasztalati alapon, automatikusan javulhatnak (Mitchell, 2006; Jordan & Mitchell, 2015). A gépi tanulási folyamatok már létező adatbázisokon alapuló modelleket építenek fel és használnak problémamegoldásra (Burkov 2019). Eszköztáruk az informatika és a matematikai statisztika metszetében helyezhető el (Mitchell 2006, Jordan & Mitchell 2015).

A ML módszerek kidolgozása már a XX. század közepén kezdetét vette (Kriesel 2007), azonban a megfelelő hardveres háttér, a nagy teljesítményű grafikus kártyával rendelkező számítógépek csak a 2010-es évektől állnak rendelkezésre (Hinton et al. 2012). A XXI. században a „big data” korát éljük, és az egyre nagyobb mennyiségű digitális adat lehetővé teszi, és meg is követeli az automatizált adatkezelést (Murphy 2012, Goodfellow et al. 2016).

A földtudományok területein, többek között a vulkanológiában és magmás geokémiában is rengeteg új adat születik a modern, digitalizált mérések során, melyek kezelése, elemzése manuálisan, csupán emberi erőforrásból kivitelezhetetlen lenne. Az adatok értelmezésének gyorsasága, pontossága és objektivitása elengedhetetlen, éppen ezért az elmúlt évtizedben robbanásszerűen megugrott a gépi tanulási módszerek vulkanológiai-geokémiai problémákra történő alkalmazásáról születő tudományos publikációk száma. A vulkánok komplex rendszerek, a megfigyelésük során gyűjtött adatok kinyerése, a releváns információk kiszűrése, a különböző eredetű adathalmazok összefésülése nem triviális feladat. Ugyanakkor ezen adatok jelentősége óriási, hiszen segíthetik a kitörések előrejelzését, a vulkáni veszély értékelését és a károk csökkentését. A folyamatos, vagy legalábbis kis időközönként végzett passzív szeizmikus, infraszónikus, geomágneses, termikus, geokémiai és felszíni deformáció mérések nagy mennyiségű adatot produkálnak, így ezek feldolgozása gépi tanulási módszereket igényel. A kiértékelt adatok információt adnak a vulkánok aktuális állapotáról, annak változásairól, melyek sok esetben a kitörések előfutárai. Számos vulkán sűrűn lakott területen fekszik, így a gyors és hatékony adatkezeléssel és értelmezéssel csökkenthetőek lehetnek a vulkánkitörések okozta nagyszabású társadalmi és gazdasági károk is. Emellett a

ML különböző módszerei és azok kombinációi hatékonyan alkalmazhatóak a geokémiai, geológiai és petrológiai „statikus” adatok vizsgálatára és osztályozására is, mint például geokémiai adatok mintázatának felismerésére, sztratigráfiai problémák megoldására, vagy akár vulkánok térbeli eloszlásának, tektonikai helyzetének értékelésére is.

A témában a közelmúltban született néhány összefoglaló tanulmány (pl.: Malfante et al. 2018, Cortés et al. 2019, Ramis et al. 2019, Carniel & Guzmán 2020), melyek jelen szakdolgozat kiindulási alapjaiként szolgálnak. A célom ezen túlmenően a főbb gépi tanulási módszerek működésének részletesebb megértése, illetve minél több szakirodalomból megismert példa összegyűjtése a ML vulkanológiai és magmás geokémiai területen történő alkalmazásáról. Az esettanulmányokból elsajátított ismeretanyag értelmében értékelem a gépi tanulás és a vulkanológiai-geokémiai adatelemzés kapcsolatát, reflektálva az adatalapú megközelítés előnyeire, illetve a gyakorlati megvalósítás esetleges nehézségeire is.

A ML elméleti hátterének és az alkalmazási lehetőségeket tárgyaló publikációknak ismerete remélhetőleg utat nyit számomra a jövőben gépi tanulási algoritmusok saját geokémiai adatokon történő alkalmazására is, mely munkám során személyes motivációt jelentett. Egy olyan innovatív területre nyertem betekintést, mely a XXI. századi tudomány és technológia világában egyre inkább magához ragadja a vezető szerepet, és ez nincs másképp a földtudományokban, geológiában sem.

2. Gépi tanulás

Dolgozatomban az angol nyelvű informatikai szakkifejezések magyar nyelvű megfelelőit az Angol-magyar informatikai szótár (Iványi 2007) fordításai szerint használom. A gépi tanulás lépéseinek és módszereinek bemutatását pedig főként olyan szakkönyvek, tanulmányok alapján végzem, melyek összegyűjtik ezeket az ismereteket, és részletes magyarázattal szolgálnak (Altrichter et al. 2006, Mitchell 2006, Sivia 2006, Hastie et al. 2009, Tan et al. 2011, Murphy 2012, Jordan & Mitchell 2015 és Goodfellow et al. 2016).

A gépi tanulás alapjai az informatikában, a matematikai statisztikában és a biológiában keresendők. Az adatok számítógépes elemzése nem új, de a hagyományos megközelítés emberek által kitalált és beállított algoritmusokat használ. A ML célja azonban, hogy egy általános kezdeti struktúrából kiindulva a számítógépek optimalizálják az algoritmusok paramétereit, a leghatékonyabb problémamegoldás érdekében (Mitchell 2006). A gépek az adatok alapján történő tanuláshoz a valószínűségelmélet eszközeit használják, melyek minden bizonytalanságot tartalmazó probléma esetén alkalmazhatóak (Murphy 2012). A gépi tanulás tulajdonképpen olyan módszerek összessége, melyek képesek automatikusan felismerni az adatmintákat, és ezek alapján, induktív módon hasonló, de ismeretlen adatokra objektív előrejelzéseket adni, bizonytalansággal terhelt döntéseket hozni. Bizonyos ML módszerek (perceptron és neurális hálók) alapjai az emberi és állati tanulási mechanizmusok, melyekhez pszichológiai és neurobiológiai megfigyelésekből merítettek ötletet, azonban a modellek matematikai technikákat alkalmazva hamar eltértek az eredeti biológiai motivációtól (Jordan & Mitchell 2015).

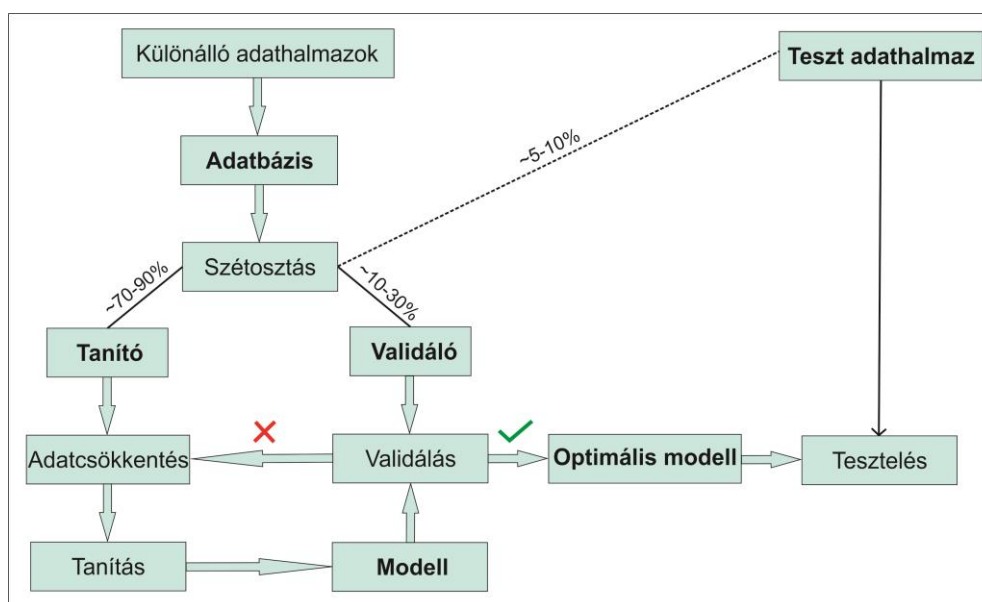
A ML módszerek egyik célja a klasszifikációs és regressziós feladatok emberi munkánál pontosabb, gyorsabb, vagy gazdaságosabb megoldása. Gépi tanulásra használható felhasználóbarát, programozási ismereteket nem igénylő, ingyenes online felület a Microsoft Azure Machine Learning Studio és a Google Cloud Machine Learning Engine. A ML területén leggyakrabban használt, saját kódok írására alkalmas programnyelvek a Python, a C/C++, a Java, a Javascript és az R (Voskoglou 2017, Molnár 2018). A vezető szerepet betöltő Python programnyelv nyílt forráskódú ML algoritmusokat tartalmazó könyvtára a Scikit-Learn (Pedregosa et al. 2011, Géron 2019), valamint a mély neurális hálók tanítására alkalmas TensorFlow és Keras (Abadi 2016, Shanmugamani 2018, Géron 2019), illetve a Pytorch (Ketkar 2017).

A továbbiakban röviden összefoglalom a gépi tanulás főbb lépéseit, valamint különböző módszereinek jellemzőit és korlátait. Célom a ML elméleti hátterének alapfokú megismerése,

ami segíti a magmás geokémiai-vulkanológiai alkalmazott szakirodalmi példák megértését, így az egyes ML algoritmusok matematikai hátterét nem tárgyalom részletesen. A fejezet megírásához a hivatkozott tudományos publikációkon kívül az ELTE TTK Fizika alapszak „deep learning” kódú „Deep learning és gépi tanulás a tudományokban” kurzusa adta számomra a megfelelő háttértudást.

2.1. A gépi tanulási folyamat lépései

Általánosságban a gépi tanulási folyamatokról elmondható, hogy a következő fő lépések építik fel: adatok szétválasztása, adatcsökkentés (lényegkiemelés), modell tanítása, modell validálása, majd tesztelése lehetőleg külső, ismeretlen adatbázison (1. ábra). Ezeket a lépéseket megelőzően, ha több különböző forrásból származó adathalmazt szeretnénk felhasználni, szükséges az adatok összefésülése.



1. ábra: A gépi tanulási folyamat lépései (Carniel & Guzmán 2020 1. ábrája és Training, Testing and Evaluating Machine Learning Models é.n. alapján)

2.1.1. Adatbázisépítés

A gépi tanulási modellekhez sok adatra van szükség, ugyanis, ha túl kevés adattal dolgozunk, akkor nem áll rendelkezésre elég információ, hogy beállítsuk a modell több ezer vagy akár több millió paraméterét. A pontos ideális adatmennyiség a megoldani kívánt probléma és a választott algoritmusok összetettségétől függ (Brownlee 2019). Abban az esetben, ha a gépi tanulási modell felépítésére olyan adatokat szeretnénk használni, melyek

több kisebb, különálló halmazt alkotnak, akkor ezeket össze kell vonnunk egy egységes nagy adatbázissá.

2.1.2. *Adathalmaz szétosztása*

Az adatokat általában két vagy három részre kell osztani, egy tanító, egy validáló (kiértékelő), illetve egy tesztelő készletre. A szétválasztás aránya attól függ, hogy mennyi adat áll rendelkezésünkre. Minél több a tanító adat, a modell annál több információt tanul meg, annál hatékonyabb lehet, viszont az adathalmazra nézve reprezentatív módon, kellő mennyiségű adattal kell kiértékelnünk a modellt (Kuhn & Johnson 2019).

Általában az adatok kb. 75%-át szokás tanításra használni, és kb. 15-20%-át validálására, abban az esetben, ha a maradék kb. 5-10%-ot a végső tesztelésre különítik el (Training, Testing and Evaluating Machine Learning Models é.n.). Amennyiben azonban lehetséges, érdemesebb a teszt adatokat egy teljesen független, külső adatbázisból meghívni, ezzel garantálva, hogy a modell hatékonysága nem az adatbázisra jellemző, szisztematikus információk bemagolásának köszönhető. Az adatszivárgás elkerülése érdekében a tanító, a kiértékelő és a tesztelő adatokat teljesen külön kell kezelni, a modell elfogultságának (bias) kiküszöbölése érdekében (Shabtai et al. 2012, Wujek et al. 2016).

Az adatok szétválasztására több eljárás létezik, például a véletlenszerű, a szisztematikus, a rétegzett és az önkényes (kényelmi) mintavételezés, a CADEX (Kennard & Stone 1969) és DUPLEX (Snee 1977) algoritmusok, valamint a próba-hiba módszer (Reitermanova 2010).

Kereszt kiértékelési (cross-validation) eljárás alkalmazása esetén nem tanító és kiértékelő készletre bontjuk az adatainkat, hanem „a rendelkezésre álló mintapontokat véletlenszerűen felosztjuk k diszjunkt részhalmazra, majd $k-1$ részhalmaz pontjait tanításra, a maradék részhalmaz pontjait pedig kiértékelésre használjuk. Az eljárást megismételjük az összes részhalmazra úgy, hogy a tanítópontok közül mindig más-más részhalmazt hagyunk ki és használunk kiértékelésre” (Altrichter et al. 2006). Ezt a validációs módszert kevés rendelkezésre álló adat esetén érdemes használni (Goodfellow et al. 2016).

2.1.3. *Adattisztítás, dimenziócsökkentés, jellemzők kinyerése*

A számítástechnika több területén, mint például az adatbányászat, a gépi tanulás, az adatminták felismerése esetében is szükséges a nyers adatok előkészítése, tisztítása, melyek a ML algoritmusok bemeneteként (input) szolgálhatnak. Az adattisztítás során a statisztikai

műveleteket érdemes a tanító és validáló halmaz szétválasztása után elvégezni, csak a tanító halmazon. Ugyanis abban az esetben, ha például a teljes adathalmaz átlagát, szórását, kvartiliseit, vagy minimum és maximum értékeit számoljuk ki, illetve a teljes adatsorban keressük a kiugró értékeket, akkor a modell a validáló halmaz statisztikai jellemzőit is megismeri, már a tanuló fázisban. Ebben az esetben a teljesítmény kiértékelése nem adna valós képet az ismeretlen adatokon történő alkalmazás hatékonyságáról (Shabtai et al. 2012). A következőkben a leggyakoribb adattisztítási feladatok megoldására sorolok fel néhány módszert a teljesség igénye nélkül.

Az ML modell tanítására való előkészületek során ki kell szűrni a hiányzó és a kiugró értékeket, zajos és hibás adatokat (Zhang et al. 2003). Ezek feltárásában segítséget nyújthat az adatpontok vizuális megjelenítése például pontdiagramon, dobozdiagramon vagy hisztogramon. A hiányzó adatokat eltávolíthatjuk, vagy pótolhatjuk a meglévő értékek átlagával vagy móduszával (Wujek et al. 2016), illetve bonyolultabb becslési eljárásokat is alkalmazhatunk, mint például a MICE (Multiple imputation by chained equations) (Royston & White, 2011). Léteznek olyan gépi tanulási algoritmusok, mint például a naiv Bayes modellek és a döntési fák, melyek jól kezelik a véletlenszerű hiányzó adatokat, nem romlik a hatékonyságuk, így nem szükséges ezen modellek tanítása előtt pótolni a hiányzó értékeket (Wujek et al. 2016).

Szükséges lehet elvégezni az adatok normalizálását vagy egységesítését, ami az adatbányászatban olyan transzformációt jelent, mellyel a különböző nagyságrendű változóinkat azonos skálára hozhatjuk, így ezeket a paramétereket kombinálni tudjuk anélkül, hogy a nagyobb értékkel rendelkezők túlsúlyba kerüljenek a modellben (Tan et al. 2011).

Amennyiben egy bizonyos tartományba vagy kategóriába eső adatok túlreprezentáltak vagy alulreprezentáltak, tehát az adatsor kiegyensúlyozatlan, az ronthatja a modell előrejelzési képességét. Ezt hisztogramon való megjelenítéssel érdemes ellenőrizni. A túlreprezentált adathalmazt véletlenszerű mintavételezéssel megfelelő számúra lehet csökkenteni (Glander 2017). Ezzel szemben alulreprezentált adathalmaz esetében alkalmazható az adatpontok véletlenszerű duplikálása, de hatékonyabb módszerek például a SMOTE (Synthetic Minority Oversampling Technique) (Chawla et al. 2002) és a ROSE programcsomag R-ben (Lunardon et al. 2014).

A nagyobb mennyiségű adat nem feltétlenül vezet jobb eredményekhez, ugyanis az irreleváns információk rontják a modellek teljesítményét. Ezért fontos azon tulajdonságok kiemelése, melyek megfelelő módon jellemzik az adathalmazt. Ezek az úgynevezett jellegzetességvektorok (feature vectors) vagy jellemzők, (Hastie et al. 2009, Tan et al, 2011).

A cél minél kevesebb számú független jellemző kinyerése, melyek azonban kellően jól leírják az adatokat (Hastie et al. 2009, Cortés et al. 2015). Ezt megtehetjük például aggregálás vagy progresszív mintavételezés segítségével (Tan et al, 2011).

Fontos elkerülni az úgynevezett „dimenzióátkot” vagy Hughes-hatást (Hughes 1968), ami azt jelenti, hogy „a dimenzió növekedésével az adatok egyre ritkábban helyezkednek el az általuk kitöltött térben” (Tan et al, 2011), ami osztályozásnál és klaszterezésnél komoly problémát jelenthet. Ennek érdekében különböző dimenziószám-csökkentő eljárások használhatóak, melyek nagy részét Delicado (2008), Hastie et al. (2009), Tan et al. (2011), Tharwat (2016) és Tharwat et al. (2017) tanulmányai alapján gyűjtöttem össze. Ide sorolható például a független komponens analízis (ICA), a nemnegatív mátrixfaktorizáció (NNMF), a szinguláris érték felbontás (SVD), a főkomponens analízis (PCA), a lineáris diszkriminancia-analízis (LDA), a klaszteranalízis (CA), a kombinált klaszter-és diszkriminancia analízis (CCDA) (Kovács et al. 2014), a faktoranalízis (FA), és a többdimenziós skálázás (MDS). A felsoroltaknál sokkal egyszerűbb módszer a korreláció vizsgálata az egyes jellemzők között, és jó korreláció esetén csak az egyik továbbvitele a ML modell bemeneteként.

A modell hatékonyságának növelése érdekében generálhatunk új jellegzetességvektorokat is, például az eredeti változók összegével, különbségével, arányával, melyekről azt gondoljuk, hogy fontos információkat hordoznak az adatainkról (Tan et al. 2011).

2.1.4. Tanítás

A gépi tanulási modellek tanítását a tanító készlet adataival végezzük. Több tanulási típust különböztetünk meg: felügyelt, felügyelet nélküli, félig felügyelt, megerősítéses és transzfer tanulást.

Felügyelt (supervised) tanulási algoritmusok alkalmazása esetén a bemeneti adatpontokat címkékkel látjuk el, melyeket a megfelelő szaktudással rendelkező kutatók, szakemberek határoznak meg. Célunk a modell kimeneti paramétereként a lehető legnagyobb valószínűséggel visszakapni a helyes címkét (Altrichter et al. 2006). A tanítás tulajdonképpen egy modell-illesztési feladat a bemeneti paraméterekhez. A paraméterekhez tartozó súlyokat a hiba metrika, vagyis a költségfüggvény (cost function, average loss) minimalizálásával lehet optimalizálni, az úgynevezett gradiens módszer segítségével, amely tulajdonképpen a legnagyobb gradiens irányába tett iterációs lépésekkel végrehajtott minimumkeresési

folyamat (Mahendru 2019). Klasszifikációs és regressziós problémákra is alkalmazható (Altrichter et al. 2006, Murphy 2012).

Felügyelet nélküli (unsupervised) tanulásnál a tanító adatpontok nincsenek címkézve, és a modell ezek elemzésével igyekszik adatmintákat keresni, például sűrűsödéseket, csoportokat. Alkalmas lehet például sűrűségfüggvény becslésre, klaszterezésre, a sokdimenziós adattérben kitüntetett „irányok” keresésére, vagy kiugró értékek szűrésére. Gyakran lényegkiemelésre, a jellegzetességvektorok kinyerésére használunk ilyen módszereket az adatelőkészítés során (Altrichter et al. 2006, Murphy 2012).

A félig felügyelt tanulás tulajdonképpen az előző két eljárás keveréke, ugyanis felcímkézett és címkék nélküli adatpontokat is tartalmaz a tanító készlete. Kiszámú, vagy drágán előállítható címkézett adat esetén jól alkalmazható. Használható klasszifikációra a felügyelt módszer tágításával, illetve klaszterezésre a felügyelet nélküli eljárás bővítésével. Ezen kívül regressziós és dimenziócsökkenési problémák megoldására is képes (Zhu et al. 2009).

A megerősítéssel tanulás során a gép folyamatosan kölcsönhat a környezetével, vagyis mindig a környezet adott állapotához igazodva igyekszik optimalizálni a modell kimentét, amely visszahat a környezetre. Ebben az esetben nem egy minta adatbázisból tanul a rendszer, hanem próbálkozások és hibák útján szerez tapasztalatot, miközben a számszerű „jutalmat” maximalizálni próbálja. Tulajdonképpen az algoritmusok feladata a megfelelő viselkedés tanulása az adott környezetben (Goodfellow et al. 2016, Sutton & Barto 2018).

A transzfer tanuláshoz akkor érdemes fordulni, ha kevés tanító adatunk van. Így egy megfelelő terjedelmű adathalmazon előtanított modellt, finomhangolást követően alkalmazhatunk a kiszámú, hasonló adatunkra (Pan & Yang 2009).

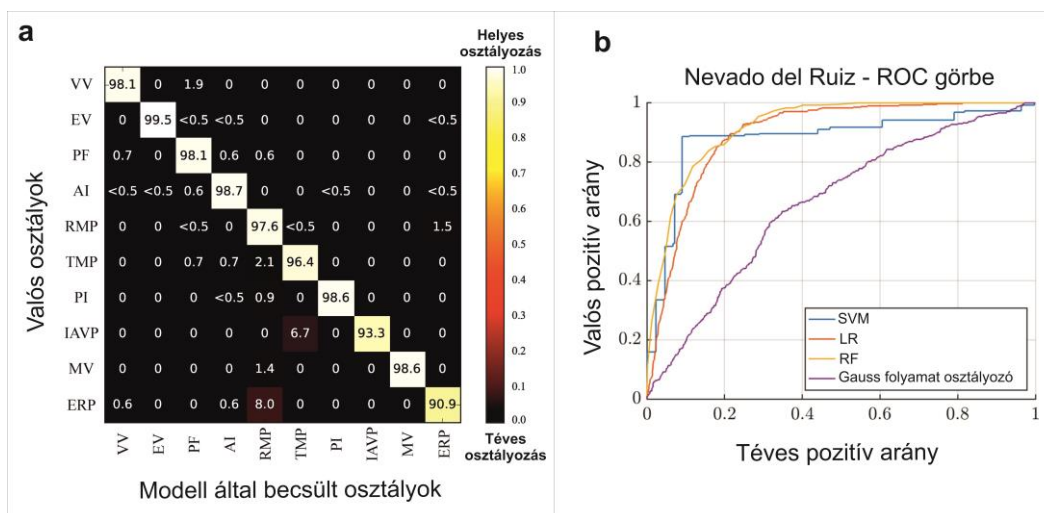
A bemutatott tanulási eljárások során alkalmazott leggyakoribb algoritmusokat, modelleket a 2.2. fejezetben ismertetem.

A tanítás során elkövetett hiba a modell alulillesztése vagy túlillesztése lehet. Előbbi esetben a modell még a tanító adathalmazra is túl gyengén illeszkedik, míg utóbbiban a modell tökéletesen illeszkedik a tanító készletre, de új adatokra gyenge becslést ad, hiszen csak „bemagolta” a tanulás során az információkat. Ezek a problémák kiküszöbölhetőek a különböző regularizációs eljárásokkal, mint például a LASSO, a Ridge regresszió és a rugalmas háló, melyekkel csökkenthető az együtthatók száma és/vagy nagysága (Wujek et al. 2016, Lever et al. 2016).

2.1.5. Validálás

A felügyelt tanulási módszereknél a modell kiértékelése során a validáló adathalmazon vizsgáljuk a modell előrejelző képességét. A modell annál jobb, minél kisebb a hibája, tehát a cél a hibát jelző metrika minimalizálása a modell hiperparamétereinek hangolásával. Hiperparaméterek például a tanulási ráta vagy a használt regularizációs technika (Training, Testing and Evaluating Machine Learning Models é.n.).

A vizsgált hiba metrika klasszifikáció esetén lehet például a valódi pozitív, a valódi negatív, a téves pozitív és a téves negatív hibából felépülő konfúziós mátrix (2.a ábra), illetve az ezek arányából számolható pontosság (accuracy), precízió (precision) és felidézés (recall) (Tan et al. 2011, Jordan 2017). A ROC görbe (receiver operating characteristic – vevő működési karakterisztika) a valódi pozitív arány (y) és a téves pozitív arány (x) függvénykapcsolatának grafikus megjelenítése. Egy jó modell a lehető legközelebb helyezkedik el az ábra bal felső sarkához, míg véletlenszerű találgatás esetén az $(x=0, y=0)$ és $(x=1, y=1)$ pontokat összekötő főátlóra illeszkedik (Tan et al. 2011) (2.b ábra). A ROC-görbe alatti terület (AUC – area under the curve) szintén alkalmas a modell kiértékelésére, ha 1, akkor a modell tökéletes, ha pedig 0,5, akkor csupán véletlenszerűen találgat (Tan et al. 2011).



2. ábra: Több osztályos konfúziós mátrix (a) és ROC görbe (b). Az a) ábrán a főátló mentén helyezkednek el a helyes becslések, míg a többi négyzetben a hibák. A rövidítésekkel jelölt osztályok olaszországi magmás provinciákra vonatkoznak (lásd: 3.2. fejezet) (Petrelli et al. 2017 5. ábrája). A b) ábra a különböző osztályozó algoritmusok hatékonyságát hasonítja össze, mely szerint az SVM, LR és RF hasonlóan jól működhetnek egy kolumbiai vulkán földrendései osztályozására (lásd: 3.1. fejezet) (Manley et al. 2020).

Regressziónál például az átlagos abszolút hiba (MAE), az átlagos négyzetes hiba (MSE), vagy ennek négyzetgyöke (RMSE), illetve a determinációs (R^2) és korrelációs (R) együttható vizsgálható (Jordan 2017, Handelman et al. 2019).

A felügyelet nélküli klaszterezés esetében végezhetünk statisztikai vagy nem statisztikai kiértékelést. A statisztikai lehet belső, vagyis felügyelet nélküli, illetve külső, vagyis felügyelt validáció. A belső kiértékeléshez tartoznak az úgynevezett particionáló és a hierarchikus eljárások. Előbbire példa a kohézió és szeparáció mértékének vizsgálata, mely azt mutatja meg, hogy az adott klaszterben mennyire közel helyezkednek el egymáshoz az elemek, illetve, hogy milyen szintű az elkülönülés a különböző klaszterek között. Mivel ezekhez a metrikákhoz nincs szükség külső információra egy külön kiértékelő adathalmazból, ezért belső indexeknek is nevezhetjük őket. A külső validációs módszerek kevés esetben használhatóak a felügyelet nélküli tanulás értékelésére, hiszen ezekhez szükség van külső információra. Ide tartozik többek között a klaszterezés pontossága, precíziója, felidézése, valamint például a klaszterek tisztasága (purity) is (Palacio-Niño & Berzal 2019).

2.1.6. *Tesztelés*

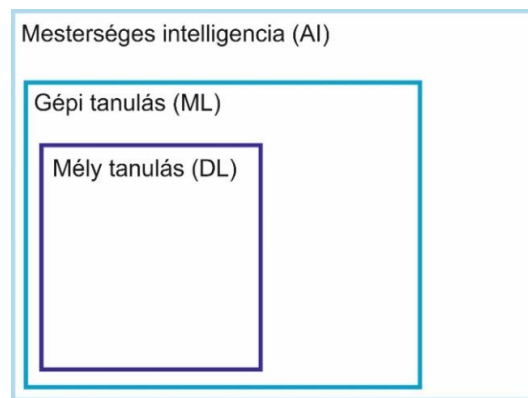
A validálás során legjobban teljesítő néhány optimális modellt az adatok szétosztása során elkülönített tesztelő készleten, vagy egy külső, teljesen független teszt adathalmazon futtatva kiválasztjuk a legkisebb hibával rendelkezőt, mely a végleges modell lesz (Jordan 2020).

A „Nincs ingyen ebéd” tétel (No free lunch theorem) (Wolpert 1996) értelmében nem létezik egyetlen univerzális modell, amely minden problémátípusra megfelelően működik, ezért sok különböző módszer fejlesztése szükséges a valóságban előforduló különféle adathalmazok lefedése érdekében. Továbbá minden modell tanítására több algoritmus közül választhatunk, tehát a gépi tanulás világa az adatok, a modellek és az algoritmusok lehetséges kombinációiból épül fel (Murphy 2012, Goodfellow et al. 2016).

2.2. A gépi tanulás főbb módszerei

A gépi tanulás, mint a dolgozatomban már korábban is említettem, a mesterséges intelligencia egyik ága, azonban maga a ML is több, egymással összekapcsolódó részterületre bontható. A klasszikus gépi tanulási módszerektől (pl.: logisztikus regresszió, klaszteranalízis) elkülönül a fejlettebb mély tanulás (deep learning, DL) (pl.: mély neurális hálók, konvolúciós neurális hálók) (3. ábra). Az előbbivel viszonylag egyszerű

függvényillesztések végezhetőek el az adatpontokon regressziós vagy klasszifikációs feladatok megoldása céljából. Az utóbbi pedig több, egymásra épülő függvényt is tartalmaz, és összetett mintázatok, például képek felismerésére is alkalmas. A klasszikustól a mély tanulásig haladva egyre több olyan komponensből épülnek fel a modellek, amelyek képesek az adatok alapján tanulni, így egyre komplexebb problémák megoldására alkalmasak, egyre inkább közelítve az emberi idegrendszer bonyolult működését (Goodfellow et al. 2016).



3. ábra: A gépi tanulás és a mély tanulás elhelyezése a mesterséges intelligencia fogalmán belül Venn-diagramon szemléltetve (Goodfellow et al. 2016 1.4. ábrája alapján)

Megkülönböztethetünk paraméteres és nem-paraméteres modelleket, attól függően, hogy a modell paramétereinek száma rögzített, vagy pedig növekszik a tanító adathalmaz kiterjesztésével. A paraméteres eljárások általában gyorsabban használhatóak, ugyanakkor nem alkalmazhatóak rugalmasan különböző adateloszlásokra. A nem-paraméteresek ezt az akadályt jól kezelik, azonban sok esetben számítási szempontból nem megoldható a nagy kiterjedésű adatbázisokon való futtatásuk (Murphy 2012). Gyakran használt paraméteres modellek például a logisztikus regresszió, a lineáris diszkriminancia-analízis, az egyszerű perceptron, a naiv Bayes osztályozók, illetve az egyszerű neurális hálók. Népszerű nem-paraméteres algoritmusok pedig a k-legközelebbi szomszéd, bizonyos döntési fák, illetve a tartóvektor-gépek. A szóban forgó paraméterek olyan konfigurációs változók a modellben, melyek értékei a bemeneti adatok alapján állapíthatók meg. Ide tartoznak például a lineáris vagy logisztikus regressziós egyenlet együtthatói, a tartóvektor-gép tartóvektorai, valamint a mesterséges neurális hálók súlyai (Murphy 2012, Brownlee 2020).

A továbbiakban a magmás geokémia-vulkanológia területén Carniel & Guzmán (2020) munkája szerint leggyakrabban alkalmazott gépi tanulási algoritmusok működését mutatom be röviden, a matematikai háttér tárgyalása nélkül. A módszerek megismerése érdekében

főként Altrichter et al. (2006), Tan et al. (2011), Murphy (2012) és Goodfellow et al. (2016) összefoglaló könyveiből, tanulmányaiból meírtettem információkat.

A **klaszteranalízis (CA)** osztályozásra alkalmazható, felügyelet nélküli módszer. A hierarchikus klaszterező algoritmus egy faszerű szerkezetet épít, ami reprezentálja a klaszterek közötti kapcsolati rendszert. A nem hierarchikus klaszterező algoritmusban pedig a meglévő csoportok összevonásával vagy szétválasztásával jönnek létre újak, bizonyos értékelési szempontok minimalizálásával vagy maximalizálásával (Tan et al. 2011) (4.a ábra). Az egyik legnehezebb feladat az ideális csoportszám kiválasztása, amelyekbe az adatokat érdemes besorolni. Ezt a számot a legtöbb esetben a kód futtatása előtt rögzíteni kell. Ennek eldöntésében segíthetnek többek között az R programnyelv NbClust csomagjába beépített technikák (pl.: elbow, silhouette, gap statistics, heuristics) (Charrad et al. 2014). A klaszterek megfelelő számának objektív megállapítására hatékonyan alkalmazható statisztikai módszer a kombinált klaszter- és diszkriminancia analízis, mellyel homogén csoportok meghatározása is lehetséges az R programnyelv ccda csomagjának alkalmazásával (Kovács et al. 2014). A módszer a hierarchikus klaszteranalízist és a lineáris diszkriminancia-analízist kapcsolja össze (Kovács et al. 2014, Kovács et al. 2015, Kovács & Eröss 2017). Normális eloszlástól eltérő eloszlású adathalmaz esetén is jól működik, azonban a kiugró értékek és a hiányzó adatpontok problémát jelentenek, így fontos az előzetes adattisztítás (Kovács et al. 2014).

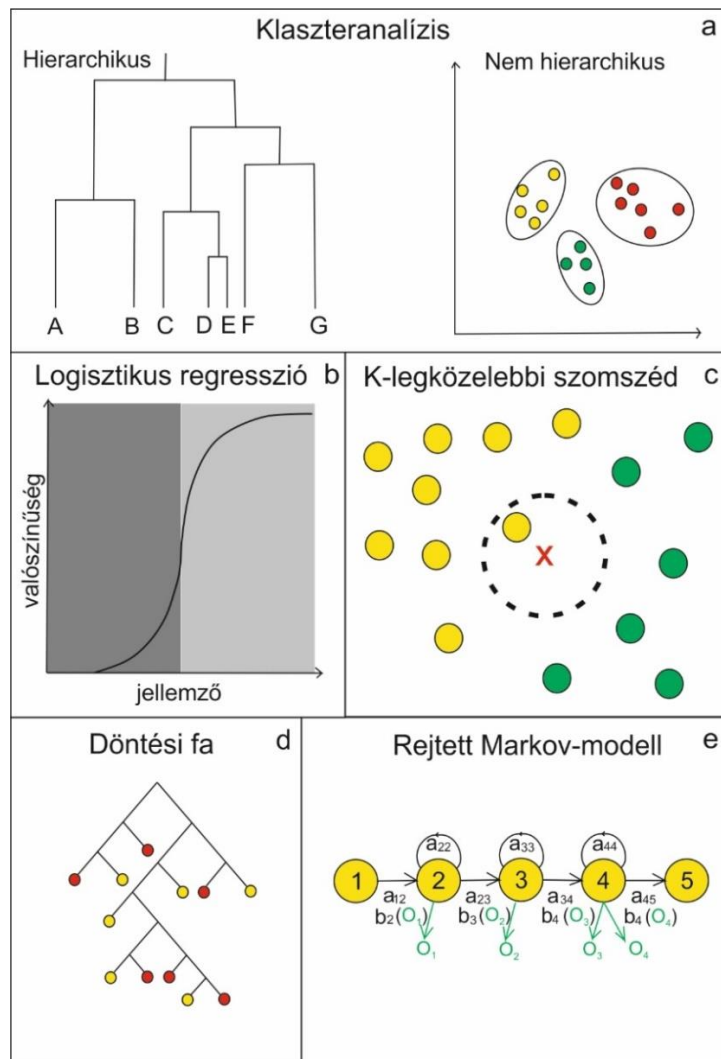
A **logisztikus regresszió (LR)** szintén klasszifikációra használható, azonban felügyelt, általánosított lineáris osztályozó modell, vagyis lineáris kapcsolat áll fenn a jellemzők és az osztályozás, mint valószínűség között. A tanuló adatokra logisztikus (szigmoid) függvény illesztésével a 0,5-nél kisebb valószínűségeket 0-hoz, a 0,5-nél nagyobbakat 1-hez rendeli (McCullagh & Nelder 1989, Murphy 2012) (4.b ábra).

A **k-legközelebbi szomszéd (KNN)** nem-paraméteres, felügyelt osztályozó, amely a bemeneti teszt adathoz legközelebb eső, tehát hozzá hasonló „k” darab tanító adat alapján dönt (Murphy 2012) (4.c ábra). Különböző nagyságrendű attribútumok esetén fontos azok előzetes skálázása, hogy a nagyobb értékekkel, és így nagyobb varianciával rendelkező jellemzők ugyanolyan szomszédsági mértékbe kerüljenek egymással, mint a kisebb értékűek és változékonyságúak (Tan et al. 2011).

A **naiv Bayes (NB)** felügyelt osztályozó algoritmusok a Bayes-is valószínűségelméleten (Bayes-tétel) alapulnak, és azt feltételezik, hogy az adott osztályhoz tartozó attribútumok egymástól függetlenek, és így becsülik az osztályhoz tartozás valószínűségét (Tan et al. 2011).

A **döntési fák (DT)** felügyelt módon, címkézett adatokkal tanítható, aciklikus, irányított gráfok, melyek minden elágazási csomópontjukban egy specifikus attribútumot hasonlítanak egy megadott küszöbértékhez, és ez alapján döntenek, hogy a jobb vagy bal ág következzen soron. Így a gyökér csúcsoktól az egymásra következő kérdéseket tartalmazó belső csúcsokon keresztül eljutunk a levél (terminális) csúcsokig, amelyekhez az osztálycímkek tartoznak (4.d ábra). A különböző algoritmusokkal tanítható döntési fák úgynevezett gyenge tanulók (weak learner), de eredményeik zsákolással (bagging) kombinálhatóak, és így robusztus modell építhető fel. A zsákolás vagy bootstrap aggregálás (bootstrap aggregating) a tanuló adatok visszatevéses mintavételezésével egymástól csak kismértékben különböző, az eredetivel azonos nagyságú adathalmazokat hoz létre, melyekkel elvégzi a tanítást, majd egyesíti az eredményeket (Tan et al. 2011). Ezen az elven alapulnak a **véletlen erdők (RF)**, melyek minden tanulási lépésben a tanulóhalmaz egy véletlenszerű részhalmazát választják ki, ezzel minimalizálva a döntési fák becslései közötti korrelációt (Breiman 2001).

A **rejtett Markov-modell (HMM)** egy olyan időbeli valószínűségi modell, melyben feltételezzük, hogy a modellezett sztochasztikus folyamatnak nincs emlékezete (Markov-folyamat). Lehetséges események sorozatát írja le, melyben minden egyes esemény valószínűsége egyedül az előző esemény során betöltött állapottól függ (4.e ábra). Az állapotok (4.e ábrán számokkal jelölve) rejtve vannak, nem láthatóak direkt módon, de megfigyeléseink (4.e ábrán „O” betűkkel jelölve) és egy adott valószínűségi eloszlás (4.e ábrán „a” és „b” betűkkel jelölve) segítségével, adott időpontokban következtetni tudunk az állapotra (Rabiner 1989, Russel & Norvig 2009). Ezt a módszert többek között a megerősítéses tanulásban, illetve időbeli mintázatok, például az emberi beszéd felismerésére használják. Utóbbi esetben a modell feladata, hogy az akusztikus megfigyelésekhez, tehát a bemeneti hanghullám jellemzőihez legnagyobb valószínűséggel tartozó állapotsort, azaz a fonémák vagy szavak sorozatát megtalálja (Rabiner 1989).

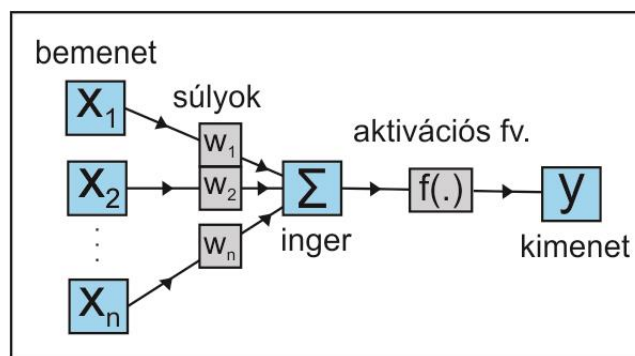


4. ábra: Gyakran alkalmazott klasszikus gépi tanulási módszerek sematikus rajzai (Carniel & Guzmán 2020 2. és 3. ábrája, valamint Tan et al. 2011 5.7. ábrája alapján). Az a)-d) ábrákon szemléltetett osztályozási módszerek az egyszerűbb klasszikus ML eljárások közé tartoznak. A c) ábra szerint az „x” jelzésű adatpont a sárga osztályba tartozik, a d) ábrán pedig a gráf elágazásai az eldöntendő kérdések, és a piros vagy sárga körök az osztálycímkek. Az e) ábrán bemutatott modell felépítése összetettebb, a számok jelzik az ismeretlen állapotokat, „a” jelöli az állapotok közötti átmenet valószínűségeit, „b” a kimenetek valószínűségeit, „O” pedig a megfigyelések sorozatát (Trujillo-Castrillón et al. 2018).

A **mesterséges neurális hálók (ANN)** „olyan, számítási feladatok megoldására létrejött párhuzamos feldolgozást végző, adaptív eszközök, melyek motivációja a biológiai rendszerektől származtatható” (Altrichter et al. 2006). Az emberi idegsejtekből (neuronokból) felépülő idegrendszer mintájára alkották meg ezeket a mesterséges számító rendszereket, olyan feladatok elvégzésére, amire a hagyományos gépi tanulási algoritmusok nem

bizonyultak kellően hatékonyak, például képfelismerési problémákra. Akárcsak az idegrendszerünkre, a mesterséges neurális hálózatokra is jellemző a párhuzamosság és redundancia, vagyis, ha néhány neuron nem működik, mások képesek átvenni a szerepüket (Altrichter et al. 2006). Bár a kezdeti kutatások során kiderült, hogy az egyszerű perceptron „csak lineárisan szeparálható osztályozási feladatok megoldására képes” (Minsky & Papert 1969), a későbbiekben az úgynevezett Hopfield-háló (Hopfield 1982), illetve a hibavisszaterjesztéses (back-propagation) algoritmus (Rumelhart et al. 1986) hatékony megoldást kínált a többrétegű perceptronok, tehát a neurális hálózatok tanításához. Az elmúlt évtizedekben egyre több matematikailag megalapozott eredmény született, és még jelenleg is intenzív kutatómunka folyik annak érdekében, hogy „egyre pontosabban meghatározzák azon feladatok körét, ahol lehet és érdemes is neurális hálózatokat alkalmazni” (Altrichter et al. 2006).

A mesterséges neurális hálók elméleti hátterét, típusait főként az Altrichter et al. (2006) által használt definíciók szerint mutatom be. Az ANN „azonos, vagy hasonló típusú – általában nagyszámú – lokális feldolgozást végző műveleti elem, neuron többnyire rendezett topológiájú, nagymértékben összekapcsolt rendszeréből áll”. A minta adatok alapján történő tanulást lehetővé tevő tanulási algoritmussal (learning algorithm) és a megtanult információ felhasználása érdekében előhívási algoritmussal (recall algorithm) rendelkeznek. A neuron „egy több-bemenetű, egy-kimenetű eszköz, amely a bemenetek és a kimenet között általában valamilyen nemlineáris leképezést valósít meg. Egy neuron rendelkezhet lokális memóriával is, amelyben akár bemeneti, akár kimeneti értékeket vagy a működés előéletére vonatkozó állapotinformációt tárolhat. A bemeneti-, vagy a bemeneti- és tárolt értékekből az aktuális kimeneti értéket egy tipikusan nemlineáris függvény alkalmazásával hozza létre, melyet aktivációs vagy transzfer függvénynek nevezünk.” Gyakori nemlineáris aktivációs függvények a lépcsősfüggvény, a tangens hiperbolikus és a logisztikus (szigmoid) függvény (Altrichter et al. 2006). A memória nélküli neuron (perceptron) bemeneti paraméterei (x_i) súlyozva (w_i) összegződnek, majd a súlyozott összegből (az ún. ingerből) a nemlineáris aktivációs függvény segítségével megkapjuk a kimenetet (választ) (5. ábra) (Rosenblatt 1958, Widrow 1960). Az ANN bemeneti, kimeneti, és köztes rejtett neuronokból épül fel, melyek általában rétegekbe rendeződnek. Megkülönböztethetünk előrecsatolt (feedforward) és visszacsatolt (recurrent, RNN) hálózatokat a neuronok kapcsolatai alapján.



5. ábra: Az egyenrangú bemenetekkel rendelkező memória nélküli neuron (perceptron) működési elve (Altrichter et al. 2006 1.2. ábrája alapján)

A **többrétegű perceptron (MLP)** „rétegekbe szervezett neuronokból áll, ahol annak biztosítására, hogy a hálózat kimenete a súlyok folytonos, differenciálható függvénye legyen, a neuronok differenciálható kimeneti nemlinearitással rendelkeznek”. Legalább három rétegből áll: egy bemeneti réteg, minimum egy rejtett réteg és egy kimeneti réteg (6.a ábra). Felügyelt osztályozó, melynek tanítása legtöbbször a hibavisszaterjesztés (back-propagation) algoritmus alkalmazásával történik. Ez az algoritmus „egy pillanatnyi gradiensen alapuló iteratív tanuló eljárás”, mely az úgynevezett lánc-szabály segítségével visszafelé számolja ki a rétegek kimeneteinél a parciális deriváltakat, és optimalizálja a súlyokat a hibafelület minimumának elérése érdekében (Altrichter et al 2006).

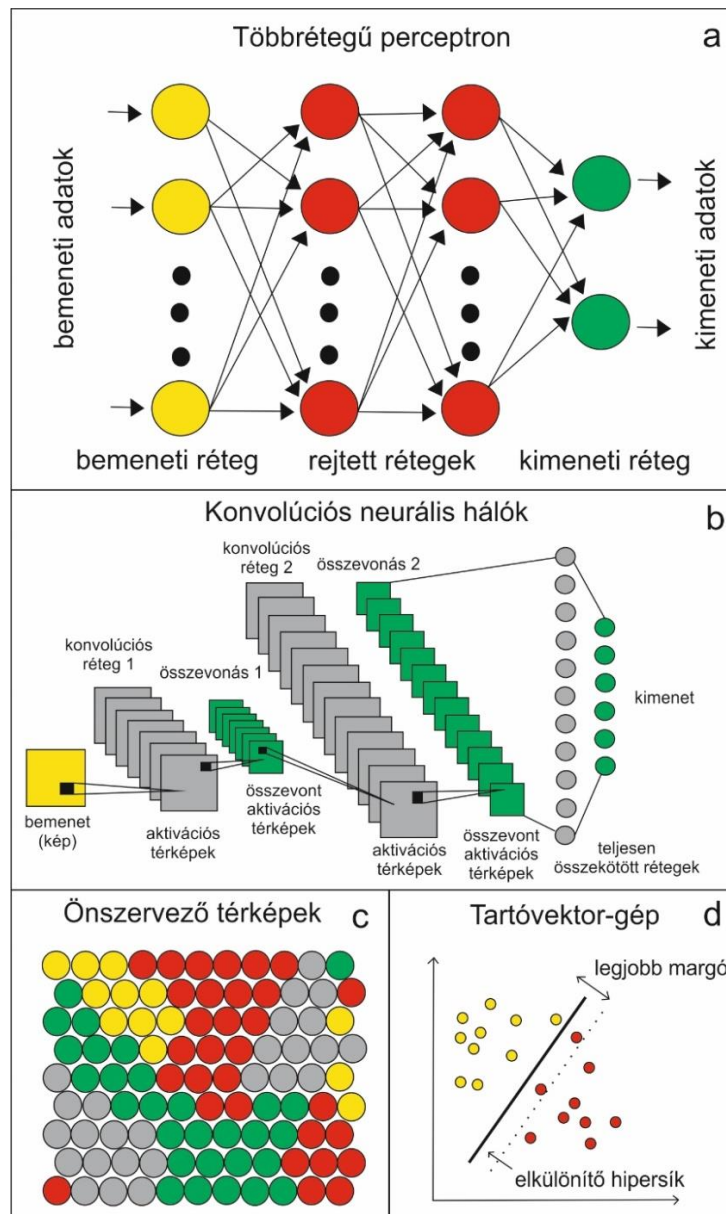
A **mély neurális hálók (DNN)** nagyszámú rejtett (mély) réteggel rendelkező összetett rendszerek, melyeket főként felügyelt módon tanítanak. A mély tanulás (deep learning) segítségével összetett problémák is megoldhatóak azáltal, hogy a modell a bonyolult összefüggéseket egyszerűbb koncepciók hierarchikus kapcsolati rendszerével reprezentálja. A DNN „mélysége” (depth) a lehető leghosszabb út a bemenettől a kimenetig, amely hosszúság azonban függ az alkalmazott algoritmusoktól, hiszen ugyanazt a műveleti sort kevesebb bonyolult lépéssel, és több egyszerűbbel is el lehet végezni (Goodfellow et al. 2016).

A **konvolúciós neurális hálók (CNN)** a DNN azon típusai, melyek legalább egy rétegükben konvolúciót alkalmaznak az általános mátrixszorzás helyett, vagyis konvolválják a bemeneti paramétereiket valamilyen szűrővel. (Goodfellow et al. 2016). Rácyszerű topológiájú adatok kezelésére alkalmasak, mint például idősorok (1D rács) vagy képek (2D rács). A konvolúciós rétegben a bemeneti adatoknak, például egy kép pixeleinek a letapogatására kernel-mátrixokat használnak, melyek különböző objektumok, élek felismeréséért felelősek, és így a konvolúciós réteg kimenete egy aktivációs térkép (feature map), vagy több szűrő esetén több aktivációs térkép lesz, melyek már nem minden egyes

eredeti pixelt, csak a képfelismerés szempontjából fontos információkat tárolják. A konvolúciót követően alkalmazott aktivációs függvény általában a ReLU (Rectified Linear Unit). A kimenet egyszerűsítését az összevonó réteg (pooling layer) végzi, például „max pooling” (Zhou és Chellappa 1988) módszerrel, mely az adott terület maximális értékét emeli ki. Az eddig leírt módszerek a CNN-ben a lényegkiemelést végzik, az osztályozás általában a konvolúciós rétegek után következő teljesen összekötött rétegeken történik (Goodfellow et al. 2016) (6.b ábra).

Az **önszervező térképek (SOM)**, vagy Kohonen-térképek (Kohonen 2001) a visszacsatolt mesterséges neurális hálók leggyakrabban alkalmazott típusai, melyek felügyelet nélküli tanulási mechanizmussal képesek a sokdimenziós bemenő adatokat két dimenzióban leképezni, és emellett kompetitív tanulással, az úgynevezett szomszédsági-függvény segítségével klaszterezni is (Tan et al. 2011) (6.c ábra).

A **Tartóvektor-gép (SVM)** (Boser et al. 1992, Cortes & Vapnik 1995, Vapnik 1998) felügyelt sztatistikai tanulás, mely alapvetően lineárisan szeparálható adatok osztályozására alkalmazható, de kiterjesztve képes nemlineáris szeparálásra és regresszióra is (Altrichter et al. 2006). A tanító adatpontok lineáris elkülönítése a sokdimenziós változó térben úgynevezett hipersíkokkal lehetséges, melyekből végtelen számú határozható meg, a cél azonban a maximális margóval rendelkező megkeresése (6.d ábra), amely valószínűleg az ismeretlen teszt adatokat is megfelelően fogja elkülöníteni (Altrichter et al. 2006, Tan et al. 2011). A módszer „a döntési határt a tanulóesetek egy részhalmazának segítségével reprezentálja, amelyeket tartóvektoroknak (support vector) nevezünk” (Tan et al. 2011). Kiválóan alkalmazható sokdimenziós adatok klasszifikációjára (Tan et al. 2011). A SVM a kernel gépek közé tartozik, melyek értelmezhetőek statikus előrecsatolt mesterséges neuronhálóként is (Altrichter et al. 2006).



6. ábra: Gyakran alkalmazott mesterséges neurális hálók működésének sematikus rajzai
(Carniel & Guzmán 2020 2. és 3. ábrája alapján)

3. Gépi tanulás a vulkanológiában és magmás geokémiában

A napjainkban alkalmazott modern, digitalizált műszeres mérések során rengeteg új adat születik a földtudományok területein, többek között a vulkanológiában és magmás geokémiában is. Gyors, gazdaságos, hatékony és objektív módon történő feldolgozásuk, kiértékelésük kihívást jelent, és megkívánja a gépi tanulási módszerek alkalmazását (Carniel & Guzmán 2020).

Az aktív vulkánok folyamatos, vagy legalábbis rendszeres műszeres megfigyelése nagy mennyiségű passzív szeizmikus, infrahang, geomágneses, termikus, geokémiai, gáz

kibocsátás és felszíni deformáció adatot produkál nap mint nap. Ezen idősorok kombinálása, illetve a kitörések előrejelzésében szerepet játszó jellemzők kiemelése nem triviális feladat, ugyanakkor óriási jelentőséggel bír a vulkáni tevékenységgel összefüggésbe hozható gazdasági és társadalmi kockázatok csökkentése érdekében.

A monitoring adatok feldolgozásán túl a „statikus” geológiai, geokémiai és kőzettani információk elemzése és osztályozása is végezhető ML modellek segítségével. Ide tartozik többek között a sztratigráfiai problémák megoldása, rétegek korrelációja például tefrakronológia segítségével, geokémiai adatsorok mintázatainak felismerése és magmás folyamatokhoz kapcsolása, vulkáni területek térbeli eloszlásának vizsgálata, valamint magmás kőzetek klasszifikálása tektonikai helyzetük szerint (Carniel & Guzmán 2020).

Mindezen vulkanológiai és magmás geokémiai kutatásokban, ahol a szakembereknek nagy adatbázisokkal kell dolgozniuk, a korábban is alkalmazott módszerek mellett egyre inkább teret hódítanak a gépi tanulási eljárások. A továbbiakban ismertetem a különböző vulkanológiai-geokémiai adattípusokat és azok jelentőségét, illetve példákat sorolok fel a gépi tanulási módszerekkel történő kiértékelésükre.

3.1. Aktív vulkánok valós idejű megfigyeléséből származó adatok kiértékelése

A vulkáni aktivitás különböző formái, mint például a piroklaszt sűrűségárák, iszapárák (laharok), hamuszórás, lávaöntés, vagy törmeléklavinák komoly gazdasági károkat okozhatnak, és az emberéletre is veszélyt jelenthetnek, különösen, mivel sok vulkán fekszik sűrűn lakott területen. Sok jelenleg is aktív vagy potenciálisan bármikor kitörést produkáló tűzhányó környezetében, a felszínen rendszeres műszeres megfigyelőrendszer működik, amely egyidejűleg több paramétert is mér. A monitoring célja a mérések alapján a vulkán működésének, a magmás folyamatoknak a megismerése és megértése. Ez lehetővé teszi a vulkáni veszély (hazard) becslését, vagyis azt, hogy egy adott területen milyen valószínűséggel következhet be pusztító hatású vulkánkitörés adott időléptékben. Az előrejelzések alapján, intézkedések foganatosításával csökkenthető a vulkáni rizikó (risk), azaz a potenciálisan veszélyes vulkáni tevékenység okozta károk bekövetkezésének valószínűsége. A vulkáni aktivitás szochasztikus előrejelzése sem egyszerű (Jaquet & Carniel 2003, Jaquet et al. 2006), de a pontos időpont, helyszín és kitörési típus determinisztikus előzetes meghatározása még bonyolultabb, sok esetben szinte lehetetlen (Carniel & Guzmán 2020). A különböző mért idősorokból a szakemberek következtetni tudnak a mélyben zajló folyamatokra, a tűzhányó aktuális állapotára, és annak változásaira, melyek kitöréshez

vezethetnek. Sok esetben ezen adatok együttes értelmezése, multiparaméteres modell alkalmazása a célravezető (Carniel et al. 2010, Jousset et al. 2012). A következőkben a vulkáni területeken mérhető különféle idősorokra térek ki részletesebben.

3.1.1. Vulkáni földrengések

A legtöbb megfigyelőrendszer által detektált, és a leghatékonyabb kitörés-előrejelző adatok a passzív szeizmikus jelek. A vulkáni földrengések a tűzhányó alatt, illetve annak közvetlen közelében pattannak ki, és a vulkáni működéssel hozhatók összefüggésbe. A kitöréseket általában a szeizmogramokon tapasztalható valamilyen változás előzi meg. A rengéseknek számos oka lehet, melyek más-más kitöréstípust és vulkáni tevékenységet jósolnak, és ezek az események a szeizmológiai adatsorokon különbözőképpen jelennek meg (McNutt 2002, Malfante et al. 2018, Ramis et al. 2019).

Megkülönböztethetjük a nagy-frekvenciájú, tehát kis hullámhosszal jellemezhető vulkanotektonikai (VT) földrengéseket. Ezeket általában töréses, nyírásos mozgások okozzák, például a növekvő magmanyomás miatt a magmakamra felső részének vetők menti elmozdulása (Lahr et al. 1994, Malfante et al. 2018). Ezzel szemben a kis frekvenciájú, hosszú periódusú (LP) rengések legtöbbször a fluidumok sekély mélységű migrációjának (Chouet 2003), illetve a magma felszín felé irányuló turbulens mozgásának következményei (Hellweg 2000, McNutt 2002). Emellett okozhatja nyomásnövekedés a magma felbuborékosodása miatt (Hellweg 2000), illetve a repedések ismétlődő szűkülése és tágulása hamu- és gázbenyomulásoknak köszönhetően (Molina et al. 2004). A robbanásos rengések kiváltója a hirtelen magma-, hamu- és gázkitöréssel járó heves vulkáni működés, melyet gyakran kísér nagyon hosszú periódusú (VLP) jel (Marchetti & Ripepe 2005). Az aktív vulkánokra jellemző szeizmikus vibráció (volcanic tremor) műszeres megfigyeléseknek köszönhetően folyamatosan mérhető (Chouet 2003, Zobin 2016). Ezt a jelenséget a magma turbulens feláramlása okozza, és szorosan összefügg az LP rengésekkel (McNutt 2002). Sok esetben jó korrelációt figyeltek meg a vibráció amplitúdója és a kitörési felhő magassága között (Zobin 2016). Az erősebb alkalmi szeizmikus vibráció a magma kigázosodására utalhat (McNutt 2002). Természetesen tektonikus vonalakhoz kapcsolódó teleszeizmikus földrengések vagy különböző robbanások hatásai, illetve antropogén eredetű szeizmikus jelek is detektálhatóak a tűzhányók közelében, melyek azonban nincsenek összefüggésben a kitörésekkel.

Az egyidejű földrengéseket a digitális szeizmométerek egyszerre észlelik, ezért fontos feladat a különböző folyamatokhoz köthető szeizmikus jelek szétválasztása, a mögöttes okok feltárásának és a vulkáni aktivitás előrejelzésének érdekében. A diszkrét eseményekhez tartozó adatok manuális elkülönítése nagyon időigényes, szubjektív, és a vulkáni krízisek, heves kitörések idején kivitelezhetetlen is lehet (Ramis et al. 2019, Cortés et al. 2019). A jelcsoportok felismerésének és osztályozásának automatizálásával azonban elérhető a szinte valós idejű kiértékelés (VSR) (Boué et al. 2016, Cortés et al. 2019, Ramis et al. 2019).

Az első fontos lépés az adatcsökkentés, azaz olyan jellemzők kinyerése, melyek a háttéresemenyről a lehető legtöbb információt hordozzák (Cortés et al. 2019). Ide tartozik például a folytonos jelsorból az eseményeket tükröző releváns változások kiválasztása (Malfante et al. 2018). A zajtól megtisztított, feldolgozott adatsorok ideális esetben függetlenek a mérőműszertől és a mérési helyszíntől (Cortés et al. 2019). Akár a jelenséget leíró fizikai modell paramétereitől is függetlenek lehetnek, így akkor is jól funkcionálnak, ha nincs tudományos konszenzus az adott szeizmikus jelet generáló tényezőkkel kapcsolatban (Cortés et al. 2019). Többek között a hullámok frekvenciája, amplitúdója, a jel időtartama, valamint a szeizmogramok auto- és keresztkorrelációs együttthatói lehetnek a ML modellek bemeneti adatai (Carniel 2014). Ezek kinyerése a gépi tanulással történő osztályozás előtt manuálisan, többek között hullámtranszformációs matematikai modelleken alapuló szűrő eljárásokkal (Langer et al. 2006, Jones et al. 2012), vagy korrelációs függvényekkel (Langer et al. 2006) végezhető el. Ezen kívül alkalmazhatóak automatizált jelkereső módszerek is, mint például az STA/LTA detektor, amely 2 lépésben keresi a szeizmikus jelet egy kiválasztott időablakban (Allen 1978, Allen 1982, Trnkoczy 2009), vagy a lineáris prediktív kódolási (LPC) algoritmusok (Del Pezzo et al. 2003). A tsfresh elnevezésű Python programcsomag is alkalmazható a jellemzők kinyerésére idősorokból (Christ et al. 2018). A ML eljárások közül a rejtett Markov-modell, egyes konvolúciós neurális hálók (Canário et al. 2020) és a visszacsatolt neurális hálók nem igényelnek előzetes adatcsökkentést, ugyanis képesek azt a klasszifikációval együtt elvégezni (Cortés et al. 2019).

Az automatizált VSR céljából alkalmazható az FFM (Failure Forcecast Method), ami egy szubjektív hiba előrejelzési módszer (Tárraga et al. 2008), de például Bayes-i megközelítéssel társítva objektív, robusztus megoldást jelenthet (Boué et al. 2015). Ugyanakkor a detektált szeizmológiai adatsorok prekonceptióktól mentes, adatalapú kiértékelésére az elmúlt kb. 20 évben inkább a különféle gépi tanulási módszereket vették igénybe. Leggyakrabban használt eljárások a rejtett Markov-modellek, a többrétegű perceptronok, a tartóvektor-gépek, a véletlen erdők, az önszervező térképek, valamint a klaszteranalízis. Ezen kívül logisztikus

regressziót, mély, visszacsatolt és konvolúciós neurális hálókat is alkalmaztak néhány esetben. A szakirodalomból összegyűjtött példákat a ML modellek alkalmazási területei, tehát a megfigyelt aktív vulkánok szerint csoportosítom Malfante et al. (2018), Cortés et al. (2019), Ramis et al. (2019) és Carniel & Guzmán (2020) összefoglaló tanulmányai alapján, az elmúlt év publikációival kiegészítve (1. táblázat).

Vulkáni terület	Alkalmazott módszer (év)
Anak Krakatau (Indonézia)	MLP (2008)
Campi Flegrei (Flegrei-mezők) (Olaszország)	MLP (2003)
Colima és Popocatepetl (Mexikó)	HMM (2009, 2018, 2019, 2020)
Deception-sziget (Antarktisz)	HMM (2007, 2019, 2020), RNN (2018)
Etna (Olaszország)	SVM (2006), HMM (2009, 2016)
Fuego (Guatemala)	DNN (2018)
Galeras (Kolumbia)	HMM (2013)
Kilauea (Hawaii)	HMM (2010)
Las Canãdas kaldera (Tenerife)	HMM (2008)
Llaima (Chile)	SVM (2014, 2016), CNN (2020)
Merapi (Jáva, Indonézia)	HMM (2001, 2006), SOM (2010)
Nevado del Huila (Kolumbia)	HMM (2018)
Piton de la Fournaise (Réunion, Fro.)	RF (2017, 2020)
Raoul-sziget (Új-Zéland)	SOM (2013)
Soufrière Hills (Montserrat)	MLP (2003), CA (2006), HMM (2012)
Stromboli (Olaszország)	MLP (1996, 2006), SOM (2006, 2008), HMM (2009)
Telica (Nicaragua) és Nevado del Ruiz (Kolumbia)	LR (2020)
Ubinas (Peru)	SVM (2018)
Vezúv (Olaszország)	MLP (2005)
Villarica (Chile)	MLP (2009), SVM (2014, 2016)
Whakaari (Új-Zéland)	RF (2020), CA (2020)

1. táblázat: Az aktív vulkánokon detektált passzív szeizmikus jelek kiértékeléséhez használt ML módszerek Malfante et al. (2018), Cortés et al. (2019), Ramis et al. (2019) és Carniel & Guzmán (2020) alapján. Az esettanulmányokat a szövegben részletezem és hivatkozom.

A feldolgozott esettanulmányok alapján megállapítható, hogy a vulkáni tevékenységhez kapcsolódó passzív szeizmikus jelek kiértékelésére kezdetben, a 2000-es évek elején főként MLP módszereket, majd a 2000-es évek második felétől egyre több esetben HMM modelleket alkalmaztak, melyek mellett helyet kapott a SOM és a SVM is. Az eredetileg főként beszédfelismerésre használt HMM szerepe még napjainkban is nagy a vulkáni földrengések

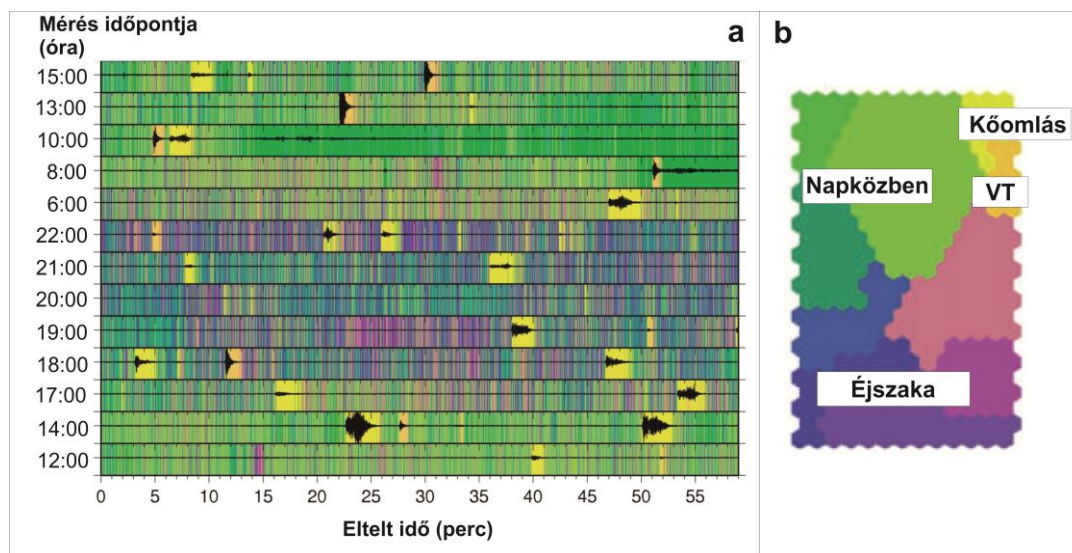
osztályozásában, de az elmúlt néhány évben már megjelentek a mély neurális hálók is (1. táblázat). A továbbiakban ismert aktív tűzhányók ML módszerek segítségével megoldott szeizmológiai adatemelési problémáit ismertetem röviden.

Az olaszországi Stromboli robbanásos kitöréseihez kapcsolódó rengéseinek osztályozására mesterséges neurális hálót (Falsaperla et al. 1996), illetve annak önszervező térképpel való kombinációját alkalmazták (Esposito et al. 2006), a háttérzaj elkülönítése érdekében pedig rejtett Markov-modellt használtak (Ibáñez et al. 2009). Az aktív kúrtók geometriája és a Stromboli-típusú kitöréseket kísérő VLP rengések közötti kapcsolat feltárására az önszervező térkép nyújtott ideális megoldást (Esposito et al. 2008). Az Etna paroxizmális lávaszökőkút-kitöréseinek előrejelzésére alkalmas módszer a HMM, amely nyugalmi, kitörés előtti, kitörési és kitörés utáni állapotokba sorolja be a szeizmológiai adatokat (Cassisi et al. 2016). Ibáñez et al. (2009) szintén HMM használatával választotta el az általános szeizmikus háttérzajt a kitörésekhez kapcsolódó vibrációtól. A lávaszökőkutak és lávaöntések elkülönítése ugyancsak a detektált szeizmikus vibráció alapján, tartóvektor-géppel valósult meg Masotti et al. 2006-os munkájában. Mesterséges neurális hálókkal különítették el a Vezúv kis magnitúdójú VT rengéseit a természetes és antropogén eredetű zajtól (Scarpetta et al. 2005). Ugyanezt a módszert alkalmazták a közeli Campi Flegrei (Flegrei-mezők) sekély mélységű VT rengéseinek és a Nápolyi-öbölben illegális halászhálóhoz kapcsolódó tenger alatti mesterséges robbantások jelalakjainak szétválasztására (Del Pezzo et al. 2003).

A vulkáni működéshez kapcsolódó földrengések és a folyamatosan detektálható zaj automatizált szétbontására a HMM is jó eredményeket biztosít, amint az a tenerifei Las Cañadas kaldera szeizmogramjainak kiértékelése során beigazolódott (Beyreuther et al. 2008).

A Merapi (Jáva, Indonézia) különböző folyamatokból származó vulkáni rengéseinek automatizált elkülönítését HMM algoritmusokkal végezték, mely főként a VT és a kőomlások által generált szeizmikus hullámok azonosítására funkcionált hibátlanul (Ohrnberger 2001, Wassermann & Ohrnberger 2001, Alasonati et al. 2006, Beyreuther et al. 2012). A Beyreuther et al. (2012) által megalkotott általános modell számára elegendő csupán egyetlen szeizmológiai állomás adatsora a hatékony osztályozáshoz. A vulkáni földrengéseken alapuló adatokon tanított HMM algoritmusok pedig más felhasználási területeken, például geotermikus erőművek környezetében észlelhető rengések értékelésében is helytállóak (Beyreuther et al. 2012). A Merapi aktivitásához köthető szeizmológiai adatoktól a napi periodicitású, emberi tevékenységből származó zaj jól elszeparálható önszervező térképpel, mely feltárja az adatok időbeli és hasonlósági kapcsolatrendszerét (7. ábra) (Köhler et al.

2010). Az Anak Krakatau (Indonézia) vulkáni tevékenysége által gerjesztett rengések kategorizálására, és a tektonikus eseményektől, valamint egyéb tranziens háttérzajtól való elválasztására jól működik a mesterséges neurális háló is (Ibs-von Seht 2008).



7. ábra: A Merapin mért rengések egy nap alatt (a), és önszervező térképpel való osztályozásuk (a, b). Az a) ábrán a szűrt szeizmogramok mögötti színezés a b) ábrán látható SOM osztályait mutatja, melyből kirajzolódik a napi emberi tevékenység okozta zaj, valamint a különböző típusú vulkáni rengések (Köhler et al. 2010 4. c) és 5. ábrája).

A hawaii Kilauea vulkáni terület Halema‘uma‘u kráterének epizodikus gázkitörései VLP rengéseket okoznak, melyek a magma transzportrendszer gáztartalmának növekedéséről informálják a szakembereket. A szeizmikus jelek azonosítását és a zajszűrést rejtett Markov-modellel végezték (Dawson et al. 2010).

A Soufrière Hills (Montserrat) vulkáni rengéseinek osztályozását a HMM nagyobb pontossággal (>97%) végezte (Hammer et al. 2012), mint a mesterséges neurális hálók (70-80%) (Langer et al. 2003, 2006). Főként a kőomlások azonosítása jelentett problémát, mivel az esemény során több különböző jelforrást kell figyelembe venni (Langer et al. 2006). A vulkán sekély magmacsatornáiban zajló magma transzportfolyamatokról árulkodó alacsony frekvenciájú rengések osztályozására Green & Neuberg (2006) klaszteranalízist alkalmazott, mely a detektált hullámformák hasonlóságai alapján klasszifikálta a szeizmogramokat.

A mexikói Colima és Popocatepetl tűzhányók földrengés idősorait főként HMM alapú osztályozó technikákkal kategorizálták és értelmezték (Cortés et al. 2009, 2019, 2020, Arámbula-Mendoza et al. 2018). Az elmúlt 2 évben alkalmazott, vulkántól független általánosított modellek képesek voltak 70-80%-os pontosságú klasszifikációra és valós idejű

kockázatértékelésre is (Cortés et al. 2019, 2020). A Telica (Nicaragua) és a Nevado del Ruiz (Kolumbia) kitöréseinek végét a vizuálisan látható jelenségeken túl a szeizmológiai adatsorok alapján is igyekeztek pontos időponthoz kötni logisztikus regresszió, SMV és RF algoritmusokkal. A Nevado del Ruiz esetén a SVM bizonyult a legjobb osztályozónak (2.b ábra) (Manley et al. 2020). Az ecuadori Cotopaxi LP és VT rengéseit 6 különböző felügyelet nélküli eljárással klaszterezték, és keresték a legalkalmasabb módszert az átfedő jelek osztályozására (Duque et al. 2020). A chilei Villarica vulkáni rengéseit MLP, illetve SVM módszerekkel értékelték ki (Curilem et al. 2009, 2014a, 2014b, 2016). A Llaima vulkán esetén szintén alkalmazták a szeizmológiai mérőállomástól független SVM modellt (Curilem et al. 2014a, 2014b, 2016), valamint a nyers adatokon direkt módon futtatható SeismicNet elnevezésű konvolúciós neurális hálót (Canário et al. 2020). A Galeras és a Nevado del Huila (Kolumbia) szeizmikus jelhullámaikat HMM segítségével osztályozták, mely vulkáni krízishelyzetben is jól működik (Bicego et al. 2013, Trujillo-Castrillón et al. 2018). A perui Ubinas esetében a SVM módszert használták (Malfante et al. 2018), míg a guatemalai Fuego adatsorán a mély neurális háló végezte az egy időben detektál különböző eredetű rengések legpontosabb klasszifikációját (Titos et al. 2018a).

A Piton de la Fournaise (Réunion) kőomlásainak és vulkanotektonikai eseményeinek azonosítása és kategorizálása véletlen erdő módszerrel történt (Hibert et al. 2017). Szintén RF-el keresték az összefüggést a szeizmikus vibráció mintázata és a kitörések dinamikája között (Ren et al. 2020).

Az új-zélandi Raoul-szigeten a szeizmikus vibrációtól az antropogén forrásokból származó zajt önszervező térképpel szeparálták. Emellett keresték a 2006-os freatikus kitöréshez köthető változásokat a szeizmogramokon (Carniel et al. 2013). A Whakaari vulkán VLP rengéseinek csoportosítását klaszteranalízissel végezték (Park et al. 2020). Ugyanitt a szeizmológiai adatsorok RF osztályozóval történő kiértékelése segítette a gőzkitörések rövid távú előrejelzését (Dempsey et al. 2020).

Az antarktiszi Deception-sziget vulkáni rengéseinek azonosítását és klasszifikációját kiválóan végezték a rejtett Markov-modellek (Benítez et al. 2007, Cortés et al. 2019; 2020).

Ahogy a felsorakoztatott példák is szemléltetik, az ideális módszer minden esetben más és más lehet, ugyanis nagymértékben függ a megoldandó probléma részleteitől. A gépi tanulási eljárások kiváló adatalapú osztályozók, a felsorolt publikációkban megalkotott modellek 70-100% közötti pontossággal teljesítettek a tesztelés során. Ugyanakkor hatékony alkalmazásuk még tartogat kihívásokat, melyek megoldása nem mindig valósítható meg egyszerűen. Ilyen például a kis mennyiségű, vagy hiányos tanító adat a felügyelt tanulási

algoritmusok esetén (pl.: Falsaperla et al. 1996, Langer et al. 2003, Ibs-von Seht 2008, Malfante et al. 2018, Titos et al. 2018a), valamint a releváns jellemzők objektív kiválasztásának nehézségei (pl.: Falsaperla et al. 1996). Mindkét akadályt jól kezelik a rejtett Markov-modellek (pl.: Alasonati et al. 2006, Beyreuther et al. 2008, 2012), utóbbit pedig az automatizált adatcsökkentést is magukban foglaló CNN (Canário et al 2020), RNN (Titos et al. 2018b) és SOM (pl.: Köhler et al. 2010) módszerek is. Magából a mögöttes vulkáni folyamat természetéből is eredhetnek a problémák, például az adott eseményhez tartozó változó vagy összetett jelek miatt (pl.: Langer et al. 2006, Esposito et al. 2006, 2008). A felügyelten és felügyelet nélkül tanuló osztályozókat összehasonlító tanulmányok a felügyelt tartóvektor-gépet (Langer et al. 2009, Lara et al. 2020, Manley et al. 2020, Murwantara 2020) tartják a legpontosabbnak. A HMM előnye ezzel szemben az, hogy folytonos, nyers szeizmogramokra, kisszámú adatra is jól alkalmazható, és lehetővé teszi a vulkántól független, valós idejű kiértékelést és online figyelmeztetést (Alasonati et al. 2006, Beyreuther et al. 2008, 2012).

Napjaink fő célja a minél több vulkán adataival jól működő, általánosított ML modellek megalkotása, és minden obszervatórium számára elérhetővé tétele. Ilyen például a REDPy (Repeating Earthquake Detector in Python) programcsomag Pythonban, melynek algoritmusával közel valós időben detektálhatóak és klaszterezhetőek az ismétlődő azonos, vagy hasonló jelalakú rengések (Hotovec-Ellis & Jeffries 2016). Szintén vulkáni rengések kiértékelése céljából hasznos Python csomagok az ObsPy (Wassermann et al. 2013) és a Msnoise (Lecocq et al. 2014). A francia vulkanológusok által fejlesztett, véletlen erdő módszeren alapuló WebObs (Beauducel et al. 2020), és a szeizmikus vibrációkat SVM és MLP osztályozókkal értékelő TREMOReC (Masotti et al. 2008) szoftvercsomagok is online elérhetőek. Szintén nyílt forráskódú a HMM alapú általánosított, valós idejű VSR feladatokat és online kockázatértékelést elvégző szoftvercsomag (pyVERSO, geoStudio, liveVSR) (Cortés et al. 2019; 2020), mely a VULCAN.ears projekt keretei között készült el. A szeizmológiai adatokon alapuló kitörés előrejelző és figyelmeztető rendszert a Horizon 2020 program részeként a 749249. Marie Skłodowska-Curie Támogatási Szerződés (<https://cordis.europa.eu/project/id/749249>) értelmében az Európai Unió finanszírozta. A felsoroltakon kívül más ingyenesen hozzáférhető szoftverek és programcsomagok is segítik az automatizált adatértékelést, és számuk, hatékonyságuk folyamatosan növekszik a kutatók magas szintű munkájának köszönhetően.

Összefoglalva, a vulkáni földrengések a kitörések legfontosabb prekursorai, és automatizált, gyors, objektív és hatékony kiértékelésük érdekében sok ML modell felépítésén

és tesztelésén dolgoznak a kutatók már mintegy 20 éve. A legjobban teljesítő, kis adatbázison is futtatható, a vulkántól és a szeizmológiai mérőállomástól független, általánosított algoritmusok ingyenes elérése a jövőben segítheti a világ minden aktív vulkánján a magas színvonalú előrejelzést és figyelmeztetést, amely súlyos károktól és veszteségektől óvhatja meg a tűzhányók környezetében élő népességet.

3.1.2. Egyéb vulkán monitoring idősorok és multiparaméteres modellek

A vulkáni földrengéseken kívül a tűzhányók felszínén, környezetében, vagy akár műholdas távérzékeléssel is detektálhatóak olyan információk, melyek alapján következtéseket lehet levonni a magmás rendszerek aktuális állapotáról, és előre lehet jelezni a valószínűsíthető kitöréseket. Ezen adatok szerepe az előrejelzésben kisebb ugyan, mint a szeizmikus jeleké, de több egyidejű adatsor alakulását együttesen figyelembe véve robusztus multiparaméteres modellek építhetők fel (Aspinall et al. 2006, Carniel & Guzmán 2020).

Infrahang adatok:

A vulkáni folyamatok, mint például a magma kigázosodása, vagy a kitörések különböző jelalakú, alacsony frekvenciájú ($\sim 0.01\text{--}20$ Hz) hanghullámokat generálnak. Ezek információkat árulnak el a mögöttes mechanizmusokról, a kitörés típusáról és erősségéről, illetve a vulkán topográfiájáról és a kráter morfológiájáról is (Matoza et al. 2018). Az eredetileg katonai célokból létesített nemzetközi monitoringrendszer (IMS) által detektált infrahang adatok lehetővé teszik a robbanásos kitörések nagy, akár több ezer km-es távolságból való észlelését. A heves robbanást kísérő akusztikus nyomáshullám tulajdonságai, mint például az amplitúdója vagy az energiája összefüggnek a kitörési oszlop magasságával. Ezzel olyan távoli tűzhányók aktivitása is megismerhető, melyek környezetében nincsenek mérőműszerek telepítve (Dabrowa et al. 2011, Matoza et al. 2018). A módszer hatékonyságát igazolja például az Etna VEI2-es erősségű explozív tevékenységének 560-1040 km-es távolságból történt monitorozása, mely összevethető az ismert vizuális jelenségekkel (Marchetti et al. 2019).

Emellett a folyamatos infrahangot kibocsátó aktív vulkánok, mint például a hawaii Kilauea (Marcillo & Johnson 2010), vagy a chilei Villarica (Johnson et al. 2012) megfigyelése egy közeli és egy távoli (több km távolságra lévő) mérőállomásról alkalmas az atmoszféra időben dinamikusan változó állapotának becslésére. A két mérési helyszín között a jel változásából megismerhetőek a szél- és hőmérsékletviszonyok, amelyeknek nem csak meteorológiai szempontból van jelentőségük, hanem a tűzhányók által kibocsátott gázok

kvantifikálását, és a kitörési felhők terjedésének modellezését is lehetővé teszik (Johnson et al. 2012).

Az infrahangon alapuló hatékony információszerzés érdekében szükséges a jelek és a forrásterületek pontos azonosítása, valamint a szél által keltett zaj, és egyéb interferáló hullámok kiszűrése (Matoza et al. 2018). A klasszikus szeizmológiai eljárások erre nem elegendők, de például a szenzorok szélvédett elhelyezése, megfelelő elrendezése megoldást jelenthet. Ezek a módszerek azonban nem alkalmazhatóak olyan mérőműszerek esetén, amelyek helyzete nem változtatható, mert például a kráterhez nagyon közeli fontos adatok mérésére specializálódtak (Carniel et al. 2014). Az ilyen jelsorok kiértékelésére alkalmazható például a nemnegatív mátrixfaktorizáció (Carniel et al. 2014), illetve több gépi tanulási algoritmus is.

A Stromboli különböző jellegű aktív periódusaihoz tartozó folytonos infraszónikus adatokat Witsil és Johnson (2020a) felügyelet nélküli k-legközelebbi szomszéd klaszterező algoritmussal osztályozta, ami a paroxizmális kitörések előrejelzésében is segítséget nyújthat. Az Etna vulkáni tevékenységéhez kapcsolódó infrahanghullámokon is alkalmaztak mintázatfelismerő klaszterező algoritmust, valamint tartóvektor-gépet, melyeknek köszönhetően egyetlen mérőállomás adatai alapján ~95%-os pontossággal meghatározhatóvá vált az éppen aktív kürtő (Cannata et al. 2011). A felügyelet nélküli ML eljárások a folytonos jelből a lényegi információ kinyerésére is eredményesen felhasználhatóak (Watson 2020). Az Etnán Watson (2020) infrahang adatok klaszteranalízisével azonosította az aktivitásban bekövetkező változást, amit egy repedés menti kitörés okozott.

A folyamatosan működő vulkánok által kibocsátott alacsony frekvenciájú hanghullámok klaszterező gépi tanulási módszerekkel megvalósuló automatizált, valós idejű, objektív kiértékelése segíti a tűzhányók viselkedésének megértését, és a paroxizmusok előrejelzését is (Watson 2020, Witsil & Johnson 2020a).

Geomágneses és elektromágneses adatok:

A vulkáni aktivitást, rengéseket gyakran kísérik a mágneses, elektromos és elektromágneses tér változásai. A robbanásos kitörések, földrengések piezomágnesességet okoznak a vulkáni kőzetekben, vagyis a hirtelen mechanikus stressz hatására bekövetkező töltésszétválás miatt néhány nT mértékű anomália keletkezik a mágneses mezőben, amely magnetométerekkel detektálható (Johnston 1997). A feláramló forró magma fűtő hatása negatív, a kitöréstermékek hűlése pedig pozitív változást idéz elő a kőzetek termoremanens mágnesezettségében (Currenti et al. 2005). A Föld elektromágneses mezejének ultra alacsony

frekvenciatartományában (ULF) észlelhető anomáliák előrejelezhetik a nagyobb vulkáni földrengéseket (Uyeda et al. 2002, Kotsarenko et al. 2005).

Például a Merapi 1990–1997 közötti, vertikális magmatranszportot, piroklaszt árákat, lávadóm növekedést majd összeomlást és izzófelhőket is magában foglaló változatos aktivitását egy hosszú periódusú (>6 év), egy közepes hosszúságú ciklusokban ismétlődő (1-2 év), és néhány hétig-hónapig tartó, epizodikusan megjelenő trendek kísérték a mágneses idősorokban. A hosszútávú mintázatnak valószínűleg termomágneses, a középtávúnak pedig piezomágneses okai lehettek, míg a néhány hetes anomáliák jól korreláltak a szeizmikus aktivitással (Zlotnicki et al. 2000).

Invreziós problémák megoldására, vagyis a felszínen mérhető mágneses anomáliák vulkanikus forrásának, például egy magmás telér paramétereinek azonosítására alkalmazhatóak mesterséges neurális hálók is (Nunnari & Del Negro 2003). Általában azonban jobban teljesítenek a legjobb paramétereket iteratív módon kereső úgynevezett genetikai algoritmusok (Goldberg & Holland 1988), melyeket az Etna 1989-es és a japán Miyake-jima 2000-es kitörése esetén használtak (Currenti et al. 2005).

Az Etnán mért, összetett folyamatok által keltett mágneses jelsor dinamikáját egy visszacsatolt SVM technikával modellezték (Jankowski et al. 2009). Szintén az Etna mágneses és gravitációs folytonos adatsorából a jellemzők kinyerését és a zajszűrést többváltozós módszerekkel végezték, melyek közül a független komponens analízis (ICA) adta a legjobb eredményt (Del Negro et al. 2008).

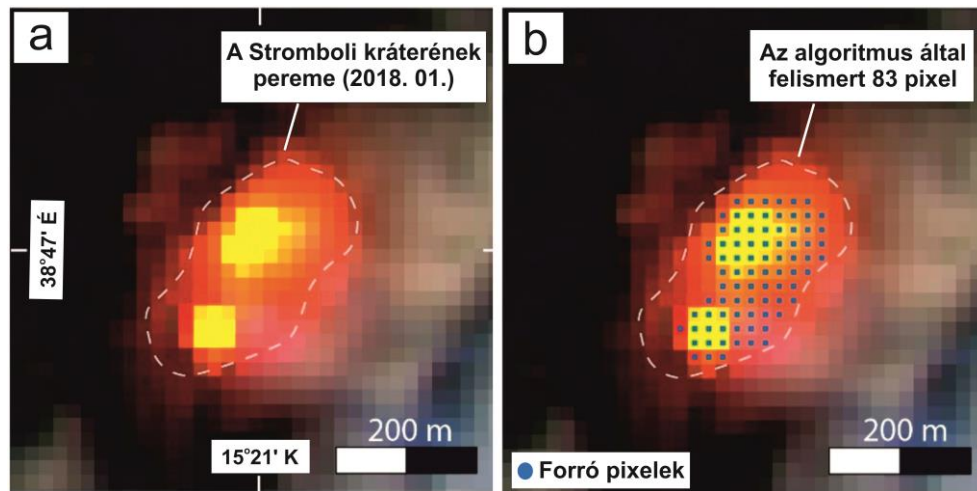
A Japán Meteorológiai Kutatóintézet által fejlesztett MaGCAP-V Windows-alapú szoftver alkalmas többek között a vulkánok mágneses adatainak elemzésére is (Fukui et al. 2013).

A vulkánokon mérhető mágneses jeleket általában nem önmagukban, hanem más idősorokkal együttesen értelmezve használják a vulkáni működés megértése és előrejelzése céljából (Fukui et al. 2013). Az inverziós problémákra, adatcsökkentésre és kiértékelésre a felsorolt néhány gépi tanulási módszer hatékony megoldást jelenthet.

Felszíni és műholdas termikus adatok:

A kráterek pereménél elhelyezett infravörös (IR) sugárzást mérő hőkamerákkal (Corradino et al. 2020, Dye & Morra 2020), illetve nagyfelbontású műholdas IR felvételek segítségével (Corradino et al. 2019, Valade et al. 2019, Coppola et al. 2020, Massimetti et al. 2020) detektálhatóak a különböző jellegű vulkánkitörések, melyek során a felszínre törő forró kőzetolvadék és vulkáni törmelék jelentős pozitív hőmérsékleti anomáliát eredményez. A

távérzékelésnek köszönhetően nehezen megközelíthető aktív tűzhányók, például forrófoltok is azonosíthatóak (Massimetti et al. 2020) (8. ábra). Az automatizált és valós idejű kiértékelés kiváló eszközei a gépi tanulási algoritmusok, amint azt a következő példák is igazolják.



8. ábra: Sentinel-2 MSI RGB felvétel a Strombolin 2018. január 3-án detektált hőmérsékleti anomáliákról (a), és a „forró” pixelek felismerése a Massimetti et al. (2020) által kifejlesztett algoritmussal (Massimetti et al. 2020 7. a) és c) ábrák).

Az Etna felszínén, a csúcsi krátereknél mért infravörös felvételeket adatcsökkentés után két különböző osztályozó alkalmazásával értelmezték. A megfigyelt terület láthatóságának mértékéről döntési fa szolgáltatott információt. A képeket pedig az úgynevezett szósák (bag-of-words) algoritmus értelmezte és csoportosította, ami a képet szöveggént kezeli, ahol a szavak a kinyert jellemzők. Ezek kombinálásával a jelenségek teljes mértékben automatikusan leírhatókká váltak, ezzel segítve a csendes időszakok, a lávaöntő, illetve a robbanásos kitörések, valamint a gáz- és hamufelhők osztályozását, és a folyamatok mélyebb megértését (Corradino et al. 2020).

Az antarktiszi Mount Erebus Stromboli-típusú kitöréseinek detektálására lehetőség nyílt szintén felszíni termikus idősorból, konvolúciós neurális háló tanításával. A Ray látató felett, a kráter peremén elhelyezett IR kamera felvételei alapján a CNN 2013 decembere és 2014 decembere között 263 kitörést azonosított. A szeizmológiai információk segítségével felcímkézett tanító adatok értelmében a modell 84%-os pontossággal teljesített. A 16%-os hibát az okozta, hogy a neurális háló azokat a kis gáz- és hamufelhőket is észlelte, amelyek az alternatívaként használt vulkáni földrengés mérésének küszöbértékét nem haladták meg (Dye & Morra 2020).

A nagy térbeli felbontású rövidhullámú infravörös (SWIR) műholdképek lehetővé teszik az olyan vulkánok működésének monitorozását is, ahol a felszíni megfigyelés nem megoldott, vagy néhány lokális hőkamera nem biztosít kellő rálátást a területre, illetve a légi megfigyelés is kockázatos. Távérzékelésből származó termikus információknak köszönhetően térképezték fel például az Etna 2017-2018-as recens lávafolyásait a Sentinel-2 műhold multispektrális műszerének (MSI) adatai alapján. Ehhez egy felügyelet nélküli klaszterező algoritmushoz folyamodtak, ami csoportokba sorolta a hasonló pixeleket. Ezután Bayes-i neurális hálóval kiválogatták ezek közül a friss lávafolyásokhoz tartozó csoportokat. A kombinált technika által közelebb kerültek a lávafolyások viselkedésének, terjedésének megismeréséhez, ami elősegíti a kockázatértékelést. Ez egy általános módszer, tehát az Etnán kívül más vulkánok esetében is hatékonyan ígérkezik (Corradino et al. 2019).

A kitörések globális mértékű megfigyelésére kifejlesztett MIROVA (Middle Infrared Observation of Volcanic Activity) érzékelési rendszer képes a forrófolt vulkanizmus automatikus felismerésére MODIS (Moderate Resolution Imaging Spectroradiometer) adatok alapján. A MIROVA projekt a világ több mint 200 tűzhányójáról tesz elérhetővé közel valós idejű termikus adatokat a www.mirovaweb.it weboldalon a Firenzei Egyetem és a Torinói Egyetem együttműködésének keretei között. Ezzel nagy mennyiségű adatot biztosítanak a kutatók számára, ami a hatékony ML algoritmusok tanításának és tesztelésének legfontosabb előfeltétele (Coppola et al. 2020).

Mind a felszíni, mind a műholdas IR felvételek értelmezésében nagy előrelépést jelentenek az elmúlt néhány évben alkalmazott képfelismerő algoritmusok, melyekhez most már a megfelelő minőségű képek, az online elérhető nagy adatbázis, valamint az informatikai háttértudás és hardveres felszereltség is adott. A termikus információ szerepe önmagában is fontos a vulkáni veszély értékelésében, de más monitoring idősorokkal együtt, multiparaméteres modellek bemenetét is képezheti (pl.: Valade et al. 2019).

Gáz kibocsátás és egyéb geokémiai monitoring adatok:

A geofizikai mérések mellett az in situ, valós idejű geokémiai megfigyeléseknek is szerepe lehet a mélyben zajló folyamatok megértésében, a kitörések előrejelzésében. Például az olasz Lipari-szigetek különböző aktivitású, vagy éppen szunnyadó tűzhányói környékén vizsgálják a fumarola-tevékenységet, a hévforrások vízkémiai összetételét, valamint a talajból történő diffúz gázszivárgást (pl.: CO₂) is. Ezáltal a mélyben kigázosodó magmáról, és a magmatározóban végbemenő folyamatokról, a termodinamikai egyensúly fennállásáról, vagy éppen felborulásáról, átalakulásáról jutnak értékes információkhoz. A hosszútávú megfigyelés

elősegíti a magmás rendszerek, mint „természetes laboratóriumok” időbeli fejlődésének, változásainak megismerését, és ezen keresztül a vulkáni veszély értékelését (Inguaggiato et al. 2018). A bazaltvulkánok paroxizmusait általában anomálishan nagymértékű CO₂ kibocsátás előzi meg, amit többek között a Stromboli esetében állapítottak meg (Aiuppa et al. 2010).

A talajból vetők, repedések mentén a radon (²²²Rn) gáz szivárgása szintén a vulkáni tevékenységhez kapcsolódik, és a vulkáni földrengések, valamint kitörések prekursora lehet, és jól korrelál más geokémiai és geofizikai paraméterekkel is (Ambrosino et al. 2020).

A kitöréseket előrejelezheti a szivárgó gázok és termálvizek izotópos összetételének megváltozása is. Például a Stromboli 2002-2003-as kitörését megelőzően a vulkán közelében sekély kutakban mintázott felszín alatti vízben a He és C elemek izotóparánya módosult. Ez összefüggésbe hozható a felemelkedő magmacsomag izotópos összetételével és a gyors emelkedés és kigázosodás közben az izotópok frakcionációjával (Capasso et al. 2005).

A Vezúv lejtőin fakadó források vízkémiai paramétereinek megváltozása kapcsolódik a szeizmikus eseményekhez, például az 1999-es földrengés alatt magmás eredetű He és CO₂ jutott az egyik vizsgált forrást tápláló vízbe. Tehát a tűzhányók környezetében a forrásvizek hidrogeokémiai vizsgálata segítheti a vulkáni aktivitás változásainak azonosítását (Federico et al. 2004).

Az aktív krátertavak vizének rendszeres vizsgálata is fontos, mivel a vízalatti freatikus és freatomagmás, heves robbanásos kitörések súlyos vulkáni veszélyt jelenthetnek. A tavak fizikai és kémiai paramétereinek változásai összefüggenek a vulkáni működés alakulásával, ezért szükséges a mérésük, illetve olyan modellek megalkotása, ami segíti a nehezen monitorozható területek értékelését is (Rouwet & Tassi 2011).

Magának a felszínre törő lávának a geokémiai elemzése, hőmérsékletének mérése is természetesen rengeteg információt biztosít a forrásmagmáról, a kitörést előidéző és megelőző folyamatokról, valamint ezek időbeliségéről. Például a hawaii Kilauea 2018-as kitörése esetén Gansecki et al. (2019) a mintagyűjtés után néhány órán belül energiadiszperzív röntgenfluoreszcens módszerrel elemezte a lávát, ami nagyban hozzájárult a valós idejű helyes kockázatértékeléshez.

A már megszilárdult kitöréstermékek, azaz a vulkáni kőzetek ásványgeokémiai jellemzői is hordozhatnak információt a várható következő kitörés idejéről, sőt akár stílusáról és erősségéről is. Például az új-zélandi Taranaki esetében a titanomagnetit ásványfázis TiO₂- és Al₂O₃-, illetve MgO-koncentrációja alkalmas a kitörést követő nyugodt időszak hosszának jóslására. A Ti, Al és Mg elemek beépülése a titanomagnetit kristályrácsába nagyban függ a kristályosodás nyomás-, hőmérséklet-, és redoxviszonyaitól. Az Al₂O₃ vagy MgO

menyiségének változékonysága a különböző magmacsomagok keveredését jelzi a mély magmatározó-rendszerben a kitörést megelőzően, és a hosszabb aktivitási szünettel korrelál. A TiO_2 koncentrációjának erős változékonysága viszont a felszín közelében történő szilárd fázisú szételegyedés következménye, és rövid nyugalmi időszakot jelez (Green et al. 2013).

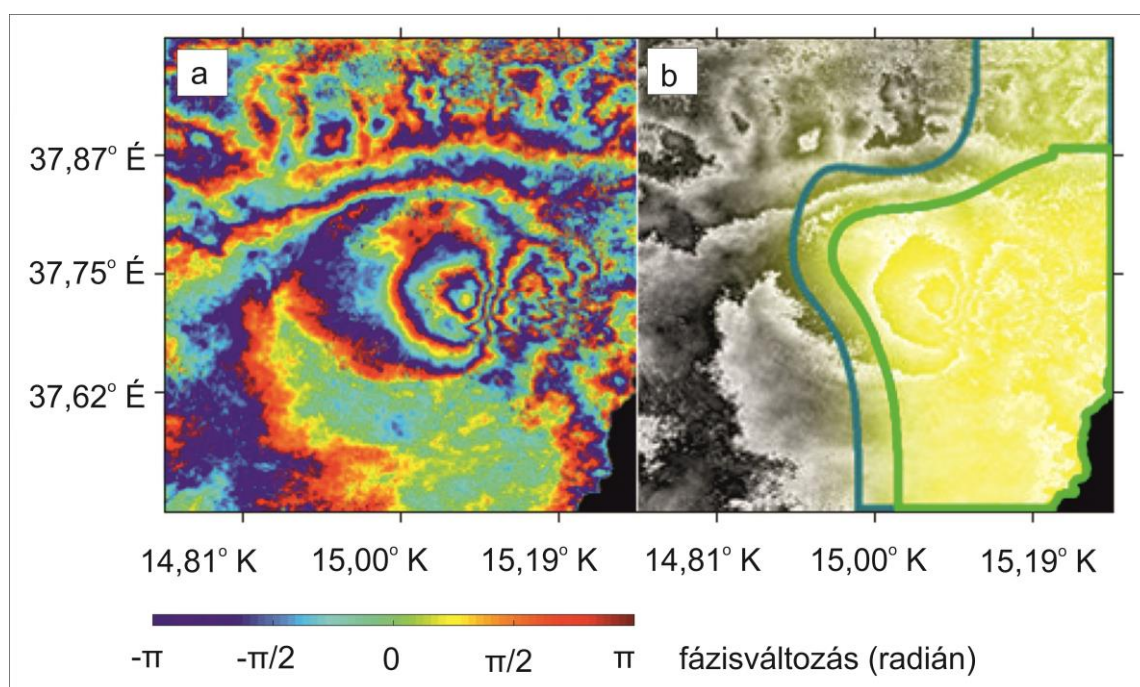
A felsorolt példák jól szemléltetik, hogy milyen sokféle geokémiai megfigyelés, mérés végezhető az aktív tűzhányók környezetében, melyek szerepet játszanak a kitörések jóslásában, a vulkáni veszély felmérésében is. A nagy geokémiai adatbázisokon taníthatóak ML algoritmusok. Például a Nápolyhoz közeli Campi Flegrei (Flegrei-mezők) területén mesterséges neurális hálót használtak a hiányos radon emisszió idősor pótlására a többi detektált geokémiai és geofizikai idősorral (pl.: szeizmikus jelek és CO_2 fluxus) való összefüggés alapján. Így sikeresen azonosították a radon anomáliákat, amik a kaldera aktivitásához kapcsolódtak (Ambrosino et al. 2020).

Nem csak a kitörést megelőző folyamatos mérések adatait kell hatékonyan kezelni, hanem például a kitörés közben detektálható vulkáni gázkibocsátást is. Erre kiváló példa a tűzhányóból az atmoszférába kerülő SO_2 felhő magasságának meghatározása műholdas UV érzékelőkkel mért adatok alapján, inverziós ML módszerekkel. A légkörbe kerülő kén-dioxid korlátozza a légiforgalmat, valamint súlyos hatásai lehetnek többek között a földi klímára is, ezért valós idejű ismerete a kockázatelemzés szempontjából nagyon fontos. Az erre a célra kifejlesztett algoritmus az adatcsökkentés, PCA és ANN eljárások kombinációja, amely gyorsan és pontosan működik (Efremenko et al. 2017). Hasonlóképp ML inverziós megoldást alkalmazott Fedkin et al. (2020) az OMI (Ozone Monitoring Instrument) UV adataira.

Távérzékeléssel észlelt felszíni deformáció és in situ videókamerás megfigyelés:

A felszín alatt megrekedt magmás intrúziók növekedése, illetve a felszínen lávadóмок épülése veszélyes vulkáni események előfutára lehet. A lávadóмок összeomlása pusztító piroklaszt árákat és piroklaszt torlóárákat generálhat (Voight & Elsworth 2000), az intrúziók pedig csökkenthetik a vulkáni felépítmény stabilitását, ami szintén jelentős kitörést eredményezhet, mint például a Mount St. Helens esetében 1980. május 18-án (Donnadieu & Merle 1998). Éppen ezért a vulkáni veszély felmérésének érdekében nagyon fontos a felszín emelkedésének megfigyelése. Az InSAR (Interferometric Synthetic Aperture Radar) technológiának köszönhetően több mint 900 vulkán felszíni deformációja detektálható cm-es pontossággal távérzékeléssel. Ezáltal olyan aktív területekre is rálátás nyílik, amelyek a felszínen nehezen érhetőek el, vagy például nem áll rendelkezésre az adott országban a megfelelő anyagi keret a megfigyelőrendszer kiépítésére (Anantrasirichai et al. 2018).

Az interferogramok értelmezésére képfelismerő CNN modelleket kezdtek alkalmazni az elmúlt néhány évben, melyek közül az úgynevezett AlexNet bizonyult a leghatékonyabbnak (pl.: Anantrasirichai et al. 2018, 2019, Gaddes et al. 2019). A neurális hálót valós és mesterségesen generált bemeneti adatok kombinációjával tanítva a modell eredményesebbé vált a zajszűrésben, és javult a felismerőképessége (Anantrasirichai et al. 2019). A Sentinel-1 InSAR adatbázis túl nagy terjedelmű ahhoz, hogy manuálisan osztályozható legyen, ezért az automatizált képfelismerő algoritmusok jelentik az egyedüli megoldást (9. ábra) (Anantrasirichai et al. 2018). Az in situ dőlésmérővel detektált deformáció pedig például a már említett japán MaGCAP-V szoftver segítségével kezelhető más paraméterekkel együtt (Fukui et al. 2013).



9. ábra: Az Etna felszíni deformációját mutató Sentinel-1 interferogram (a), valamint a CNN model (AlexNet) segítségével becsült deformáció valószínűsége (b). A b) ábrán a sötétzöld kontúron belül a deformáció valószínűsége $>0,5$ és a világoszöldön belül pedig $>0,8$ (Anantrasirichai et al. 2018 8. c) ábrája).

A tűzhányók videókamerás megfigyelése is egy lehetséges megoldás a vulkáni működés rögzítésére, azonban krízishelyzetben lehetetlen az események nyomonkövetése, a felvételek értelmezése emberi erőből. Szükséges tehát egy gyors, objektív módszer, mint például a Villarica 2015-ös aktivitását 93%-os pontossággal osztályozó ANN, de valószínűleg alternatívát jelenthetnek más algoritmusok is, mint például a SVM, a DT, vagy különböző CA technikák (Witsil & Johnson 2020b).

Több idősort együttesen értelmező modellek:

A Föld mintegy 1500 aktív vulkánjának legtöbbje csak távérzékelés segítségével monitorozható. A Sentinel műholdcsalád szabadon hozzáférhető adatai nagyban hozzájárulnak ehhez, és így kiépülhetett a MOUNTS (Monitoring Unrest from Space) megfigyelőrendszer, melynek célja a vulkáni tevékenység globális mértékű észlelése. Műholdas deformáció, SO₂ kibocsátás és termikus anomália méréseket kombinálnak felszíni szeizmológiai információkkal. A kiértékeléshez ML algoritmusokat használnak (CNN), ezzel segítve a Föld vulkanizmusának valós idejű megismerését és összetett megértését. Több recens vulkáni tevékenység esetében is megfelelően működött a MOUNTS rendszer, mint például az etiópai Erta Ale 2017-es, a guatemalai Fuego, a hawaii Kilauea és az indonéz Anak Krakatau 2018-as, illetve a Réunion szigeti Piton de la Fournaise 2018–2019 kitörése (Valade et al. 2019).

Nem csak globális skálán, hanem egyes vulkánokra külön-külön is kidolgozhatóak multiparaméteres modellek, melyek többféle idősoron alapulnak. Például az Etna geofizikai és geokémiai monitoring adatainak kombinációjával tanítva a döntési fák bizonyultak a legmegfelelőbb eljárásnak a vulkáni tevékenység értékelésére és a paroxizmusok előrejelzésére (Hajian et al. 2019).

A vulkáni veszély figyelmeztetési fokozatának eldöntésében HMM nyújthat segítséget, amely a MULTIMO (Multi-disciplinary monitoring, modelling and forecasting of volcanic hazard) projektnek köszönhetően szabad hozzáférésű geofizikai adatok (Carniel et al. 2006) felhasználásával határozza meg a vulkánok állapotát és az azok közötti átmenet valószínűségét (Aspinall et al. 2006).

A mérési technológiák fejlődése és az automatizált kiértékelési módszerek elterjedése mellett még mindig sok kihívást tartogat a kitörések előrejelzése. A fizikai és kémiai modellek, valamint az adatalapú ML algoritmusok csak valószínűségi értékeket rendelnek az események bekövetkezéséhez, a monitoring információk felhasználásával. A társadalom biztonsága érdekében azonban a cél a vulkáni működés pontos előrejelzése és a vulkáni rizikó csökkentése, melyre a probabilisztikus módszerek nem minden esetben nyújtanak kielégítő megoldást. Ráadásul ahhoz, hogy a felügyelt gépi tanulás működhessen, szükséges a nagy tanító adatbázisok pontos előzetes ismerete, a kutatók precíz munkája. A felügyelet nélküli algoritmusok szerencsére előismeretek nélkül, nyers adatokon is futtathatóak, azonban eredményüket célszerű fizikai vagy éppen termodinamikai kontextusban is validálni. Mivel a ML nem a klasszikus természettudományos törvényszerűségek, egyenletek szerint jár el,

hanem csupán az adatokban keres mintázatokat, ezért akár fizikailag lehetetlen előrejelzést is adhat. Egyes algoritmusok, főképp a mély neurális hálók olyan összetettek és bonyolultak, hogy eredményeiket esetleg nehezen fogadja el a társadalom, illetve az sem egyértelmű, hogy a ML hibájából származó téves előrejelzésért kit terhel a jogi felelősség. A gépi tanuláshoz tehát vannak korlátai a vulkanológiai-geokémiai használatban, de az adatok hozzáférhetőségének, mennyiségének és minőségének javulásával, a modellek fejlődésével egyre nagyobb szerepet kap, és a köztudatba is fokozatosan kezd beépülni, elfogadottá válni. Ez a viszonylag új megközelítés sokban megkönnyíti az automatizált, gyors, hatékony, gazdaságos és előítéletektől mentes adatelemzést, amit a modern, digitális műszeres mérési technológiák meg is követelnek (Poland & Anderson 2020).

3.2. „Statikus” magmás geokémiai adatok kiértékelése

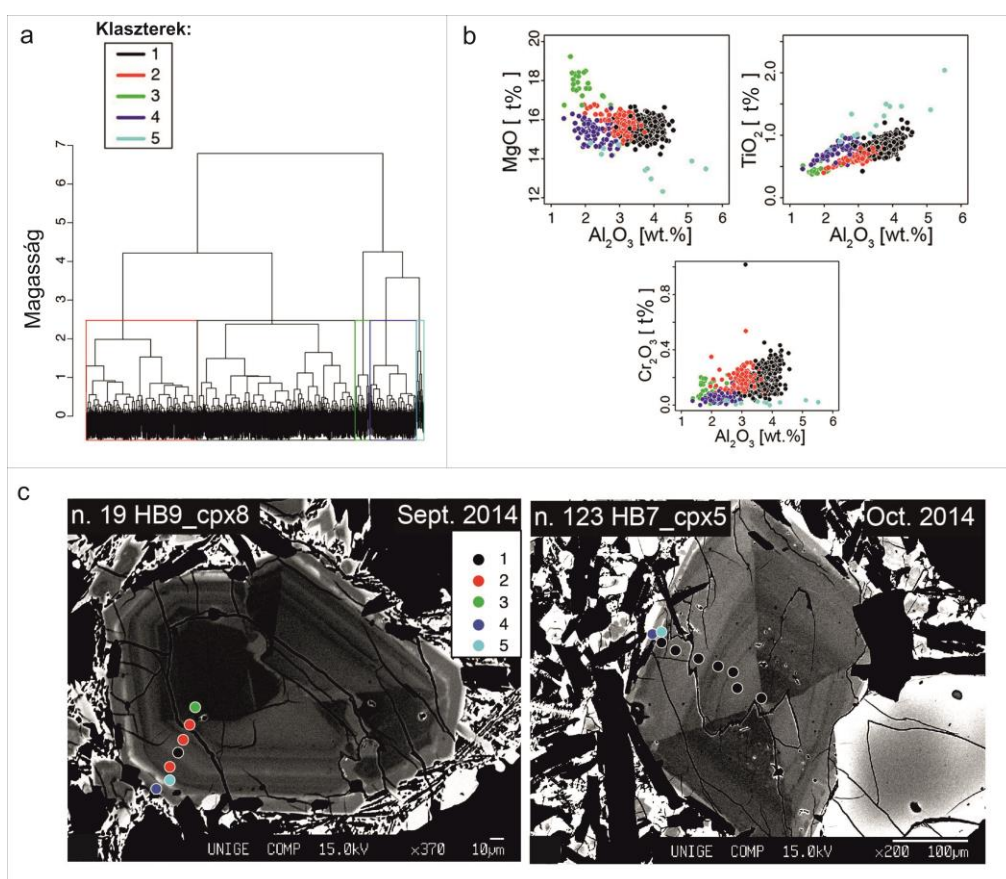
A gépi tanulási módszerek olyan adatok kezelésére is felhasználhatóak, melyek nem a vulkánok valós idejű megfigyeléséből származnak, hanem például korábbi vulkánkitörések termékeinek geokémiai elemzéséből. Az így keletkező nagy adatbázisok automatizált kiértékelésével olyan geokémiai, geokronológiai, sztratigráfiai és egyéb problémák kezelhetők hatékonyan, melyeket manuálisan nehézkes, vagy akár lehetetlen lenne megoldani (Carniel & Guzmán 2020). A továbbiakban olyan eseteket mutatok be, melyek a magmás geokémia legfontosabb megoldandó problémaköréhez tartoznak.

Magmás folyamatok feltárása geokémiai adatok alapján:

A kitöréstermékek, tehát a vulkáni kőzeteket felépítő ásványfázisok és az egykori kőzetolvadékot reprezentáló kőzetüveg, valamint az ásványokban elhelyezkedő olvadék- és fluidzárványok kémiai összetétele és annak változásai azoknak a folyamatoknak az üjjlenyomatai, amelyek a magmatározóban a kitörést megelőzően, a magmacsatornában a kitörés közben, vagy akár a kőzet utólagos mállása során lezajlottak. A főelem- és nyomelem-koncentrációk meghatározása nagyműszeres méréseket igényel, melyek aránylag lassú és költséges folyamatok, ráadásul gondos mintaelőkészítést követelnek meg. Mindazonáltal, a geokémiai adatoknak köszönhetően betekintést nyerhetünk azokba a mechanizmusokba, amik elvezettek a vulkánkitörésekhez.

A magmás geokémiai nagy adathalmazok gépi tanulási módszerekkel történő kiértékelésére legtöbb esetben különböző CA módszereket használnak (Carniel & Guzmán 2020). Például a Stromboli vulkán 26–13,8 ka közötti kitörési periódusából származó ásványok kristálykémiai adatait hierarchikus CA eljárással osztályozták, ami elősegítette az

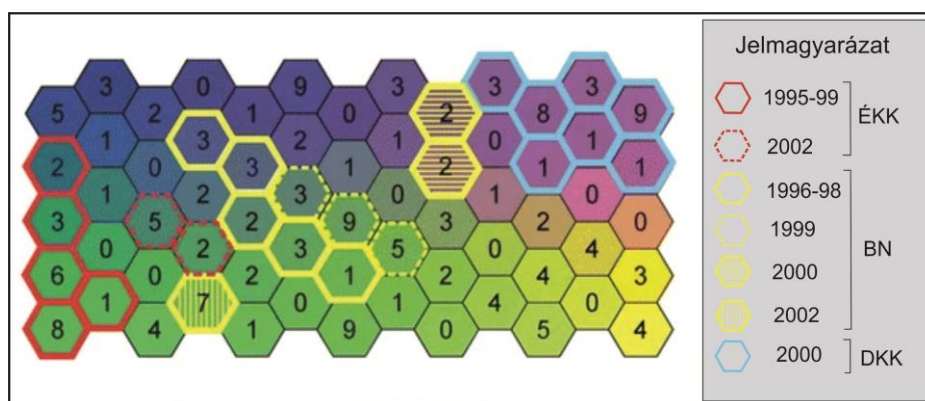
egykori magmakeveredési folyamatok feltárását (Cortés et al. 2007). PCA és CA alkalmazásával csoportosította főelem-koncentrációjuk alapján a Virunga-hegység (Kelet-Afrika) különböző térbeli elhelyezkedésű vulkáni kőzeteit Barette et al. (2017). Caricchi et al. (2020) szintén hierarchikus CA segítségével osztályozta az izlandi Holuhraun-Barðarbunga 2014-2015-ös kitöréséből származó klinopiroxén kristályok kémiai zónásságát (10. ábra). A főelemek és a nyomelemek koncentrációja alapján is sikerült elkülöníteni a szektorzónás és oszcillációs zónás területeket, valamint a külső reakciószegélyeket, melyek mind más-más magmás folyamat, például forró kőzetolvadékkal való interakció, hűlés, vagy eltérő kemizmusú magma hozzákeveredésének következményei (Caricchi et al. 2020).



10. ábra: Az a) ábra az izlandi Holuhraun-Barðarbunga 2014-2015-ös kitöréséből származó klinopiroxén kristályok kémiai zónásságának hierarchikus klaszteranalízisét szemlélteti, a b) ábra pedig a meghatározott klaszterekre jellemző főelem összetételt. A c) ábra az egyes klaszterekhez tartozó zónásság BSE képeit mutatja (Caricchi et al. 2020 4-6. ábrái alapján)

A klaszteranalízisen kívül többször használt felügyelet nélküli osztályozó a SOM is, ami hozzájárult például az Etna 1995-2005 közötti kitöréstermékeinek klasszifikálásához fő- és nyomelemek alapján. Az eredményeknek köszönhetően elkülöníthetővé váltak a különböző

magmacsatornákhöz tartozó egyidejű kitörések, valamint kirajzolódtak az ugyanazon kürtőhöz kapcsolódó időbeli trendek a kémiai összetételben (11. ábra). Továbbá az eredmények értelmében sikerült elkülöníteni legalább két eltérő geometriájú sekély mélységű magmakamrát a magmatározó-rendszer részeként, melyekben eltérő kémiai egyensúlyi állapot uralkodott (Corsaro et al. 2013). Ugyancsak önszervező térképpel dolgozták fel a Vezúv, a Campi Flegrei (Flegrei-mezők) és Ischia teljes kőzet főelem- és nyomelem-koncentráció értékeit, illetve a különböző elemarányokat. A kapott eredményeket hagyományos módszerekkel, izotóparányok segítségével validálták (Esposito et al. 2021).



11. ábra: Az Etna 1995-2005 közötti kitöréstermékeinek osztályozása fő- és nyomelemek alapján önszervező térképpel. A hatszögek jelzik az azonos csoportba tartozó adatmintákat, a számok a csoport elemeinek számát. Hasonló színek tartoznak a hasonló geokémiai tulajdonságokhoz, így kirajzolódnak a különböző aktív kürtőkhöz tartozó kitöréstermékek. A kráterekhez köthető időbeli trendeket, vagyis a magma fejlődését a piros, sárga és kék kontúrok jelölik az Északkeleti Kráter (ÉKK), a Bocca Nuova kráter (BN) és a Délkeleti Kráter (DKK) esetén (Corsaro et al. 2013 4. a) ábrája).

Felügyelet nélküli ANN technika is alkalmazható a kőzettípusok automatizált csoportosítására főelem-koncentrációjuk szerint. Vulkáni szigetívek bazalt, andezit, dácit és riolit kőzetmintáinak azonosítását 76-100%-os pontossággal végezte Lacassie et al. (2006).

A felügyelt és felügyelet nélküli ML technikák alkalmasak lehetnek a nagy kalderaformáló robbanásos kitörésekhez, illetve a kevésbé heves és a lávaöntő tevékenységhez tartozó geokémiai adatok elkülönítésére, és az átmenetet irányító folyamatok feltárására (López et al. 2021).

A geokémiai adatok feldolgozását és vizuális megjelenítését segíti például a GCDkit csomag R programnyelvben (Janoušek et al. 2006), valamint az újonnan kifejlesztett pyrolite Python programcsomag is (Williams et al. 2020). Fontos megjegyezni, hogy a kőzetek,

ásványfázisok, zárványok és közetüveg főelem- és nyomelem-összetételének gépi tanulással való feldolgozása mellett szükség van a terepi munka, a közettani megfigyelések és a klasszikus geokémiai megközelítés kombinálására a magmás folyamatok mély megismerése és megértése érdekében (Carniel és Guzmán 2020).

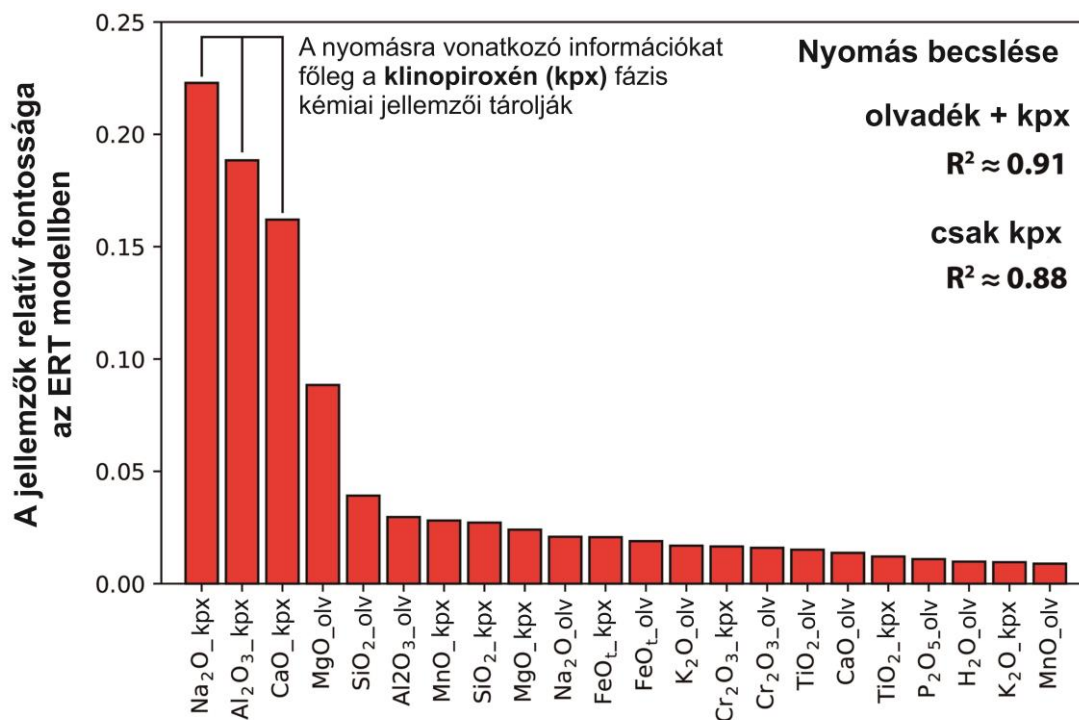
Adataalapú geotermobarometria:

A magmatározó fizikokémiai paramétereinek változásai összefüggenek a vulkáni aktivitás alakulásával, tehát többek között a kitörés előtti nyomás- és hőmérsékletviszonyok feltárása kulcsfontosságú a kitöréshez vezető folyamatok megismerésének érdekében. Az ásványfázisok és a mélyben őket körülvevő közetolvadék között fennálló kémiai egyensúlyi állapot esetén, termodinamikai modellek segítségével kiszámolható a kristályosodási nyomás és hőmérséklet. Ebből következtetések vonhatóak le a magmatározó mélységéről, illetve arról, hogy mi okozhatta a kitörést. Ezen kívül bemenetei lehetnek további modelleknek, mint például a diffúziós kronometria, mely a magmás folyamatok időbeliségét adja meg (Costa et al. 2020), és elengedhetetlen hozzá a kristályosodási hőmérséklet pontos meghatározása.

A hagyományos geotermobarometriai egyenletek termodinamikai törvényszerűségeken alapulnak, éppen ezért csak szigorú korlátok között alkalmazhatóak. Specifikus összetételű olvadék, valamint leszűkített nyomás- és hőmérséklet tartomány esetén jól működnek a klasszikus modellek. Azonban általánosan, az olvadék kémiai összetételétől, illetve a tektonikai helyzettől függetlenül alkalmazható modell csak a gépi tanulás segítségével alkotható meg. Az adat alapú eljárások nem helyettesíthetik a hagyományos egyenleteket, de kiegészítik őket, kibővítik a felhasználási lehetőségeiket. A ML preconcepciók nélkül, nem előre definiált termodinamikai reakciók alapján oldja meg a problémát, hanem kizárólag az adatokból kiindulva. Így olyan kémiai elemek összetételének mintázatát is figyelembe veheti, amelyek a fizika és kémia szabályai szerint nem kötődnek direkt módon a nyomás és hőmérséklet változásához, de a ML modell előrejelzésének hatékonyságát mégis jelentős mértékben befolyásolják (Petrelli et al. 2020).

Például klinopiroxén-olvadék kémiai összetétel adatokon ML algoritmusokat tanítva meg tudták határozni a kristályosodási nyomást és hőmérsékletet szubalkáli és alkáli magmás rendszerek esetén is. Ezáltal összehasonlíthatóvá vált a különböző kemizmusú és tektonikai helyzetű magmatározók kitörés előtti állapota. Ráadásul az alkalmazott gépi tanulási folyamat során a nyomás becslésében csupán a klinopiroxén fázis elemkoncentrációi játszottak fontos szerepet (12. ábra). Ennek köszönhetően lehetőség nyílt az olvadék tulajdonságaitól és a hőmérséklettől független klinopiroxén barometria modell megalkotására. Az új, adat alapú

becslési módszert az izlandi Barðarbunga magmás rendszerén alkalmazták, ami segítette a korábban túlbecsült magmatározó mélységek korrigálását (Petrelli et al. 2020).



12. ábra: A véletlen erdőhöz hasonló ERT (extremely randomized trees) módszer szerint a nyomás becsléséhez a klinopiroxén ásványfázis (kpx) kémiai jellemzői lényegesen fontosabbak, mint a közetolvadéké (olv), ezért megalkotható egy csak kpx adatokra épülő geobarométer (Petrelli et al. 2020 7. a) ábrája)

A zónás ásványszemcsék kémiai összetételének fluktuációja nem csak a kristállyal egyensúlyban lévő közetolvadék fizikai és kémiai paramétereinek megváltozásából ered, hanem szerepet játszik az egyes kémiai elemek diffúziója is. Az diffúzióból származó hatás csökkentésének, szűrésének érdekében szintén alkalmazhatóak ML algoritmusok (Musu et al. 2021).

Összességében a kitörés előtti kristályosodási nyomás és hőmérséklet becslésének korlátait jelentős mértékben tágítja az adatalapú, automatizált gépi tanulási módszerek használata, azonban mellettük továbbra is szükség van a klasszikus termodinamikai egyenletekre a ML modellek tanítása, validálása és kalibrációja érdekében. Fontos megjegyezni azt is, hogy a ML algoritmusok csak nagy adatbázison tanítva jelentenek robusztus megoldást, a nagy felbontású ásvány- és közetüveg-geokémiai adatok előállításához azonban időigényes és költséges műszeres mérések szükségesek (Petrelli et al. 2020).

Azonos korú és eredetű vulkáni törmelékes rétegek korrelációja:

A szórt vulkáni anyag, illetve más vulkáni törmelékes üledékek korának, geokémiai jellemzőinek meghatározása, a kitörési központ azonosítása lehetővé teszi többek között a rétegek korrelációját, a vulkánkitörések és ülepedési események időbeliségének megismerését (Brandmeier & Wörner 2016, Petrelli et al. 2016). Több gépi tanulási módszer is segítséget jelenthet ezen problémák megoldásában.

Például Petrelli et al. (2017) SVM algoritmust alkalmazott az olaszországi pliocén-pleisztocén magmatizmus során keletkezett tefra rétegek korrelációjára, >90%-os pontossággal (2.a ábra). Az online elérhető GEOROC adatbázisból (Walker et al. 2005, Sarbas 2008) 17 fő- és nyomelem-koncentráció értékei szolgáltatták a modell bemeneti adatait. A különböző vulkáni területek elkülönítésén túl a szórt vulkáni üledék eredetét is sikerült azonosítani. Az eredmények értelmében korrigálni kellett a Római Magmás Provincia (Közép-Olaszország) vulkáni tevékenységének korábban hibásan meghatározott idejét (0,8 Ma helyett 1,4 Ma) (Petrelli et al. 2017).

Tefrakronológia céljából ANN és RF eljárások kombinációját is hatékonyan használta Bolton et al. (2019) az alaszakai holocén vulkanizmus folyamatainak feltárására. Pignatelli & Piochi (2021) pedig Nápoly környéki tefra rétegek Sr és Nd izotóparányokat is tartalmazó geokémiai adatai alapján igyekezett azonosítani a kitörési központokat különböző ML módszerekkel.

Cirkon nyomelem-koncentráció értékek és U-Pb koradatok többváltozós diszkriminancia-analízissel és ML algoritmussal történő osztályozása segíti a Pannon-medence miocén korú, Si-gazdag, szórt vulkáni rétegeit létrehozó magmás folyamatok megértését, lehetővé téve a tefra- és piroklasztit sztratigráfia helyes alkalmazását a térségben (Lukács et al. 2021a, 2021b).

A piroklaszt árák üledékei is összehasonlíthatóak geokémiai és geokronológiai adataik alapján, például CA és LDA együttes használatával. A többváltozós statisztikai módszereknek köszönhetően rövid idő alatt pontos megoldást találtak a Középső-Andok (Dél-Amerika) részét képező nagy térfogatú neogén ignimbrit rétegek térbeli és időbeli mintázatainak feltárására (Brandmeier & Wörner).

A gépi tanulási technikák gyorsan és hatékonyan működnek, valamint képesek nagy adatbázisok kezelésére, azonban az összetett geokronológiai és sztratigráfiai problémák megoldásában nem helyettesíthetik a kutatók szaktudását és precíz munkáját, csak a segítségükre lehetnek (Bolton et al. 2019).

Magmás kőzetek tektonikai helyzetének értékelése:

A magmás kőzetek eredeti, képződési tektonikai helyzetének feltárása kulcsfontosságú a komplex területek tektonikai, geológiai fejlődésének rekonstruálásához. A kollíziós és szubdukciós zónákhoz, valamint a távolodó lemezszegélyekhez eltérő magmás folyamatok, és ennek következtében különböző kemizmusú kőzetolvadékok társulnak. Korábban az óceáni hátságokhoz, óceáni szigetekhez, illetve szigetívekhez tartozó bazaltokat egyes nyomelemek koncentrációi, arányai alapján különítették el, például Ti-Zr-Sr diagramokon ábrázolva, vagy az Y/Nb érték segítségével (Pearce & Cann 1973). Később a nagy, teljes adatbázisok megjelenése lehetővé tette a vizsgálatok kiterjesztését n-dimenziós adattérre is (Snow 2006). Minél több kémiai paramétert vesz figyelembe a modell, annál pontosabb becslés várható (Verma & Díaz-González 2020). Tehát a főelem- és nyomelem-koncentrációkat, valamint a Pb, Sr és Nd elemek izotóparányait együttesen figyelembe véve a magmás kőzetek hozzárendelhetőek különböző tektonikai környezetekhez (Petrelli & Perugini 2016). A szabad hozzáférésű és nagy mennyiségű geokémiai adatokon különböző ML osztályozók taníthatóak. A GEOROC (Walker et al. 2005, Sarbas 2008) és a PetDB (Walker et al. 2005) online adathalmazok alapján a SVM osztályozó átlagosan 93%-os pontossággal csoportosította a kőzeteket Petrelli & Perugini (2016) szerint.

Ueki et al. (2018) SVM, RF és egy többváltozós regressziós módszert használt 8 különböző tektonikai környezethez tartozó geokémiai adatsor elkülönítésére. Az egyes csoportok meghatározásához legalább 17 koncentráció értéket és izotóparányt kell figyelembe venni. Ilyen nagy mennyiségű, magas dimenziószámú adat esetén a leggyorsabb és leghatékonyabb megoldást azok a ML algoritmusok nyújtják, melyek képesek a sokdimenziós adattérben klaszterezni. Ugyanakkor a modellek tanításához minél több, előzetesen precízen meghatározott adatra van szükség, ami a kutatók összetett kőzettani-geokémiai-geofizikai munkáját igényli (Petrelli & Perugini 2016, Ueki et al. 2018).

A felsorolt példák alapján megállapítható, hogy a ML módszerek megjelenése forradalmasította a nagy geokémiai adatbázisok kezelését, gyors, automatizált és objektív megoldást kínálva. A magmás rendszerek azonban összetettek és bonyolultak, termodinamikai és fizikai törvényszerűségek, tektonikai folyamatok határozzák meg a működésüket. Éppen ezért az adatalapú algoritmusok csak körültekintően, nagy tudású szakemberek felülvizsgálata mellett alkalmazhatóak annak érdekében, hogy a természet törvényeinek ellentmondó, a valóságtól elrugaszkodott eredmények semmi esetre se váljanak elfogadottá. A gépi tanulás tehát sok esetben megkönnyítheti a kutatók munkáját, de fontos a

modellek által adott becsléseket kritikusan szemlélni, és a klasszikus módszerek eredményeivel is összehasonlítani (Petrelli & Perugini 2016, Carniel & Guzmán 2020).

Az egyre nagyobb mennyiségű és jobb minőségű szabadon hozzáférhető geokémiai adatnak köszönhetően az elmúlt években elindult a megbízható, robusztus ML modellek fejlesztése. Olyan internetes platformok létrehozásával, ahol a már kidolgozott és tesztelt módszerek elérhetőek, ez az új megközelítés beépülhet a vulkanológiai-geokémiai gondolkodásmódba, elősegítve minél több újszerű, de konzisztens tudományos eredmény születését.

4. A gépi tanulás értékelése magmás geokémiai-vulkanológiai szemszögből

A gépi tanulási technikák alkalmazása rengeteg új lehetőséget nyitott meg a magmás geokémia és vulkanológia terén. A gyors, objektív, hatékony, automaizált adatelemzés lehetővé teszi a nagy mennyiségű monitoring adat közel valós idejű kiértékelését, a vulkáni veszély felmérését, a potenciális kitörések által okozható társadalmi és gazdasági károk csökkentését. Emellett a „statikus” nagy adatbázisok kezelésében, kiértékelésében, a mögöttes folyamatok feltárásában is nagy segítséget nyújthatnak a ML módszerek. Alkalmazásukhoz nagy mennyiségű, jó minőségű bemeneti adatra, valamint megfelelő számítógépes kapacitásra van szükség.

A feldolgozott szakirodalmi példák alapján a szeizmológiai adatok kiértékelése főként SOM, SVM, MLP, illetve a beszédfelismerés alkalmazási területéről átvett HMM módszer segítségével történt, az infrahang jelek kezelését pedig felügyelet nélküli osztályozók segítették. A geomágneses adatsorok értelmezésére változatos, előrecsatolt és visszacsatolt módszereket használtak. A termikus információkat hordozó IR felvételeket és a műholdas felszíni deformáció méréseket képfelismerő CNN algoritmusokkal csoportosították. A gáz kibocsátás idősorokhoz ANN technikát, míg a „statikus” geokémiai adatok mintázatának feltárására felügyelet nélküli eljárásokat (pl.: CA, SOM) alkalmaztak. Tefrakronológia és a magmás kőzetek tektonikai helyzetének értékelése céljából pedig a felügyelt osztályozók (pl.: SVM, ANN, RF) bizonyultak a leghatékonyabbnak.

Az összegyűjtött esettanulmányok bizonyítják, hogy nincs egyetlen jó megoldás, nincs egy univerzális modell, hanem minden problémára más ML algoritmusok adják a legjobb becsléseket, valamint minden eljárás használatának vannak előnyei és hátrányai. Sok esetben nem feltétlenül korrelál a modell összetettsége a hatékonysággal, több klasszifikációs probléma megoldásához, például szeizmogramok osztályozására elegendő lehet a kis

számítógépes kapacitást igénylő, egyszerűbb gépi tanulási algoritmusokhoz folyamodni (pl.: CA, MLP). Természetesen a képfelismerés, vagy más bonyolult adatmintázatok elemzése megköveteli a mély neurális hálók bevetését. Így nem csak a korábban példákkal szemléltetett műholdas felvételek osztályozása, hanem más komplex, manuálisan rengeteg időt igénylő munkamenet is automatizálható lehetne a magmás geokémia-vulkanológia tudományterületén.

Ilyen például a geokémiai osztályozáshoz társított ásványszöveti polarizációs mikroszkópos felvételek, pásztázó elektronmikroszkópos visszaszórt elektronképek (BSE) elemzése (pl.: Kahl et al. 2015, Jankovics et al. 2015, 2019). A mikroszkópos, kétdimenziós képeken képelemző szoftvereket is alkalmazhatunk, mint például az ImageJ (Rasband 1997), ugyanakkor a kristályok, például a plagioklász szövetének pontos megismerése érdekében a zónásság térbeli, háromdimenziós jellemzőit is figyelembe kell venni (Cheng et al. 2017, Cheng & Costa 2019). A különböző irányú kétdimenziós metszetek feldolgozására megoldást jelenthetnének a CNN modellek, azonban tanításukhoz nagy mennyiségű, helyesen felcímkézett képre lenne szükség, ami viszont ugyancsak nehézkes és hosszadalmas emberi munkát és szakértelmet igényelne.

A kutatók, szakemberek ismeretei, tudása természetesen nem pótolható gépekkel, azonban segíthető, kiegészíthető, gyorsítható. A felügyelt tanításhoz szükséges a tanító halmaz előzetes manuális osztályozása, illetve fizikai, kémiai, termodinamikai törvényszerűségeket követő eljárások előzetes alkalmazása. A ML algoritmusok által adott predikciót is érdemes a klasszikus természettudományos elveken nyugvó eredményekkel összevetni. Az adatalapú becslés ugyanis a természetben megvalósulhatatlan kimenetet is adhat, ezért fontos kritikusan szemlélni.

A gépi tanulási algoritmusok adatalapú megközelítésből származó, statisztikailag helytálló, de természettudományos szempontból esetleg hibás eredményei, és az ebből származó téves, vagy éppen elmaradt riasztások, vulkánkitörés előrejelzések etikai és jogi kérdéseket vetnek fel. Emellett a társadalom és a tudományos kutatók egy része is szkeptikusan viszonyul ehhez a gyorsan fejlődő, újszerű, modern adatelemző megközelítéshez. A kutatói köztudatba való beépülést, a ML használatának népszerűsítését segítheti a vulkanológia-magmás geokémiai adatokon tanított, majd tesztelt, jól teljesítő modellek online, szabadon elérhetővé tétele. A legtöbb általam feldolgozott, ML módszereket alkalmazó tudományos publikáció megadja a kész kód internetes elérhetőségét is.

A ML modellek közzététele azonban még nem vonja automatikusan maga után a széleskörű alkalmazás lehetőségét. Természetesen az új problémákhoz kalibrálni,

finomhangolni kell a korábban megalkotott algoritmusokat. A geokémiai adatok esetén még így sem egyszerű felhasználni a már rendelkezésre álló kódokat. A magmás geokémiai adatok előállítása egyrészt költséges és időigényes nagyműszeres mérésekkel történik. A másik jelentős probléma a műszerkezelő szoftverek változatossága, a kinyerhető nyers adatsorok formai különbözőségei, melyekből egységes, megfelelő minőségű adatbázist kell építeni a ML modellek tanítása előtt. Ezt sokan manuálisan, például Microsoft Excel felületen végzik. Ez a meglehetősen lassú munkafolyamat jelentős mértékben gyorsítható programozási ismeretek segítségével, többek között Python vagy R programnyelven. Az adatbázisépítésre létrehozott univerzális kód a későbbiekben hozzáadódó, friss adatokon is egyszerűen futtatható, ezáltal lényegesen megkönnyítve, felgyorsítva az adatelőkészítést. Én is bekapcsolódtam egy ilyen munkafolyamatba, Kiglics Mátyás, ELTE IK programtervező informatikus mesterszakos hallgatóval. A lézerablációs induktív csatolású plazma tömegspektrométerhez (LA-ICP-MS) tartozó Iolite kezelőszoftverből származó cirkon kristály nyomelem-koncentráció adatokkal foglalkozunk. Az adatrendezés során többek között összefésülhetők az eltérő formájú, elrendezésű táblázatok, kiszűrhetők a kimutatási határ alatti mérési eredmények, valamint elvégezhetők bizonyos korrekciós műveletek. A programozásnak köszönhetően mindezt csak egyszer, egy univerzális kód formájában kell megírni, amit aztán a későbbiekben új nyers adatokon is futtathatunk, illetve szükség szerint átalakíthatjuk, kibővíthetjük további problémák megoldása céljából.

Dolgozatomban igyekeztem felvázolni a vulkanológia és magmás geokémia leglényegesebb adattípusait, mérési módszereit, valamint a kiértékelési problémákat. Az összegyűjtött alkalmazási példák nagy száma kiválóan szemlélteti, hogy milyen gyorsan fejlődő, innovatív, produktív tudományterület a ML modellekkel történő adatelemzés. A felsorolt példákon kívül még néhány további, vulkanológiához kapcsolódó feladat megoldását is segíti a gépi tanulás. Használható például vulkáni területeken a földcsuszamlások és egyéb gravitációs tömegmozgások megindulására való hajlam modellezésére (Pham et al. 2020, Si et al. 2020), valamint a vulkánok belső szerkezetét feltérképező müontomográf mérések kiértékelésére is (Carbone et al. 2014, Yang et al. 2018).

Az általam feldolgozott téma a geokémia, a geofizika, a vulkanológia, az informatika és a matematika határterületén helyezkedik el. Ennek köszönhetően rengeteg számomra új fogalmat, folyamatot, módszert, összefüggést ismerhettem meg. Nagy öröömre szolgál, hogy betekintést nyerhettem a gépi tanulás modern szemléletmódjába, valamint a vulkanológiai-geokémiai adatokon való alkalmazás lehetőségeibe és korlátaiba, és az így szerzett ismereteket a jövőben szeretném gyakorlati munkában is kamatoztatni.

5. Összefoglalás

Jelen szakdolgozat egy modern, innovatív adatelemzési megközelítést, a gépi tanulást, valamint annak alkalmazási lehetőségeit hivatott bemutatni a magmás geokémia és vulknaológia területén.

A vulkánok valós idejű megfigyeléséből származó nagy adathalmazok gyors és pontos kiértékelése segítheti a magmás folyamatok feltárását, a vulkáni veszély felmérését és a kitörések előrejelzését. A korábbi kitöréstermékek közettani bélyegeinek és kémiai összetételének részletes vizsgálata pedig egyéb vulkanológiai, geokronológiai, sztratigráfiai és tektonikai kérdésekre adhat választ. Az elmúlt mintegy 20 évben erre a célra egyre nagyobb arányban használták az automatizált, objektív, idő- és költséghatékony gépi tanulást. A releváns szakirodalomból példákat gyűjtöttem a vulkán monitoring idősorok, valamint egyéb magmás geokémiai információk gépi tanulási feldolgozására. Továbbá értékeltem az adatalapú eljárások előnyeit és korlátait a klasszikus fizikai-kémiai modellekkel szemben.

A megismert esettanulmányok alapján megállapítottam, hogy nincsen egyetlen univerzális algoritmus, amely képes lenne minden probléma hatékony megoldására, hanem a választott gépi tanulási módszert mindig a vizsgált adatok tulajdonságaihoz és a megválaszolandó kérdés összetettségéhez szükséges igazítani. A számítógépes technológia fejlődésének és a szabadon elérhető adatbázisok bővülésének köszönhetően a lehetséges megoldások egyre szélesebb halmazából lehet meríteni.

Erről a gyorsan fejlődő vulkanológiai-geokémiai adatfeldolgozási munkamódszerről az elmúlt években már születtek naprakész, összefoglaló nemzetközi tanulmányok, de a téma hazai szakirodalma még hiányos. Céлом egy magyar nyelvű, alapos áttekintés elkészítése volt, amely segítheti a matematika, informika és geológia ezen határterületének megértését, rávilágít a benne rejlő lehetőségekre, és ösztönzi a gyakorlati felhasználást.

6. Köszönetnyilvánítás

Köszönet témavezetőmnek, Dr. Haranginé Lukács Rékának a szakmai támogatásért és lelkesítésért, valamint Dr. Csabai István konzulensnek a matematikai-informatikai szakszöveg ellenőrzéséért. Köszönet Dr. Bondár Istvánnak és Kalmár Dánielnek a szeizmológiai szakszöveg átnézéséért. Köszönet illeti Maurizio Petrellit, valamint a „DL és ML a tudományokban” kurzus oktatóit a ML bemutatásáért. Kiglics Mátyást, a Python programozás tanításáért, valamint Dénes Dorottyát a dolgozatíráshoz adott technikai segítségért. Hálás vagyok Oelberg Ottónak a sok türelemért, szakmai és személyes támogatásért.

7. Irodalomjegyzék

- Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G. & Isard, M. 2016: Tensorflow: A system for large-scale machine learning. —: 12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16), pp. 265-283.
- Aiuppa, A., Burton, M., Caltabiano, T., Giudice, G., Guerrieri, S., Liuzzo, M., Murè, F. & Salerno, G. 2010: Unusually large magmatic CO₂ gas emissions prior to a basaltic paroxysm. — *Geophysical Research Letters* **37/17**.
- Alasonati, P., Wassermann, J., Ohrnberger, M., Mader, H., Connor, C. & Coles, S. 2006: Signal classification by wavelet-based hidden Markov models: application to seismic signals of volcanic origin. — *Statistics in Volcanology* **1**, 161-174.
- Allen, R. 1982: Automatic phase pickers: Their present use and future prospects. — *Bulletin of the Seismological Society of America* **72/6B**, S225-S242.
- Allen, R. V. 1978: Automatic earthquake recognition and timing from single traces. — *Bulletin of the Seismological Society of America* **68/5**, 1521-1532.
- Altrichter M., Horváth G., Pataki B., Strausz Gy., Takács G., Valyon J. 2006: Neurális hálózatok — Panem Kft.
- Ambrosino, F., Sabbarese, C., Roca, V., Giudicepietro, F. & Chiodini, G. 2020: Analysis of 7-years Radon time series at Campi Flegrei area (Naples, Italy) using artificial neural network method. — *Applied Radiation and Isotopes* **163**, 109239.
- Anantrasirichai, N., Biggs, J., Albino, F. & Bull, D. 2019: A deep learning approach to detecting volcano deformation from satellite imagery using synthetic datasets. — *Remote Sensing of Environment* **230**, 111179.
- Anantrasirichai, N., Biggs, J., Albino, F., Hill, P. & Bull, D. 2018: Application of machine learning to classification of volcanic deformation in routinely generated InSAR data. — *Journal of Geophysical Research: Solid Earth* **123/8**, 6592-6606.
- Arámbula-Mendoza, R., Reyes-Dávila, G., Dulce, M. V.-B., González-Amezcuca, M., Navarro-Ochoa, C., Martínez-Fierros, A. & Ramírez-Vázquez, A. 2018: Seismic monitoring of effusive-explosive activity and large lava dome collapses during 2013–2015 at Volcán de Colima, Mexico. — *Journal of Volcanology and Geothermal Research* **351**, 75-88.
- Aspinall, W., Carniel, R., Jaquet, O., Woo, G. & Hincks, T. 2006: Using hidden multi-state Markov models with multi-parameter volcanic data to provide empirical evidence for alert level decision-support. — *Journal of Volcanology and Geothermal Research* **153/1-2**, 112-124.
- Barette, F., Poppe, S., Smets, B., Benbakkar, M. & Kervyn, M. 2017: Spatial variation of volcanic rock geochemistry in the Virunga Volcanic Province: Statistical analysis of an integrated database. — *Journal of African Earth Sciences* **134**, 888-903.
- Beauducel, F., Lafon, D., Béguin, X., Saurel, J.-M., Bosson, A., Mallarino, D., Boissier, P., Brunet, C., Lemarchand, A. & Anténor-Habazac, C. 2020: WebObs: The volcano observatories missing link between research and real-time monitoring. — *Front. Earth Sci.*
- Benítez, M. C., Ramírez, J., Segura, J. C., Ibanez, J. M., Almendros, J., García-Yeguas, A. & Cortes, G. 2006: Continuous HMM-based seismic-event classification at Deception Island, Antarctica. — *IEEE Transactions on Geoscience and remote sensing* **45/1**, 138-146.
- Beyreuther, M., Carniel, R. & Wassermann, J. 2008: Continuous hidden Markov models: application to automatic earthquake detection and classification at Las Cañadas caldera, Tenerife. — *Journal of Volcanology and Geothermal Research* **176/4**, 513-518.
- Beyreuther, M., Hammer, C., Wassermann, J., Ohrnberger, M. & Megies, T. 2012: Constructing a Hidden Markov Model based earthquake detector: application to induced seismicity. — *Geophysical Journal International* **189/1**, 602-610.
- Bicego, M., Acosta-Munoz, C. & Orozco-Alzate, M. 2012: Classification of seismic volcanic signals using hidden-Markov-model-based generative embeddings. — *IEEE transactions on geoscience and remote sensing* **51/6**, 3400-3409.

- Bolton, M. S., Jensen, B. J., Wallace, K., Praet, N., Fortin, D., Kaufman, D. & De Batist, M. 2020: Machine learning classifiers for attributing tephra to source volcanoes: an evaluation of methods for Alaska tephtras. — *Journal of Quaternary Science* **35/1-2**, 81-92.
- Boser, B. E., Guyon, I. M. & Vapnik, V. N. 1992: A training algorithm for optimal margin classifiers. — Proceedings of the fifth annual workshop on Computational learning theory, pp. 144-152.
- Boué, A. S., Lesage, P., Cortés, G., Valette, B., Reyes-Dávila, G., Arámbula-Mendoza, R. & Budi-Santoso, A. 2016: Performance of the ‘material Failure Forecast Method’ in real-time situations: A Bayesian approach applied on effusive and explosive eruptions. — *Journal of Volcanology and Geothermal Research* **327**, 622-633.
- Brandmeier, M. & Wörner, G. 2016: Compositional variations of ignimbrite magmas in the Central Andes over the past 26 Ma—A multivariate statistical perspective. — *Lithos* **262**, 713-728.
- Breiman, L. 2001: Random forests. — *Machine learning* **45/1**, 5-32.
- Burkov, A. 2019: The Hundred-Page Machine Learning Book by Andriy Burkov. — *Expert Systems* **5/2**, 132-150.
- Canário, J. P., Mello, R., Curilem, M., Huenupan, F. & Rios, R. 2020: In-depth comparison of deep artificial neural network architectures on seismic events classification. — *Journal of Volcanology and Geothermal Research* **401**, 106881.
- Cannata, A., Montalto, P., Aliotta, M., Cassisi, C., Pulvirenti, A., Privitera, E. & Patanè, D. 2011: Clustering and classification of infrasonic events at Mount Etna using pattern recognition techniques. — *Geophysical Journal International* **185/1**, 253-264.
- Capasso, G., Carapezza, M., Federico, C., Inguaggiato, S. & Rizzo, A. 2005: Geochemical monitoring of the 2002–2003 eruption at Stromboli volcano (Italy): precursory changes in the carbon and helium isotopic composition of fumarole gases and thermal waters. — *Bulletin of volcanology* **68/2**, 118-134.
- Carbone, D., Gibert, D., Marteau, J., Diamant, M., Zuccarello, L. & Galichet, E. 2014: An experiment of muon radiography at Mt Etna (Italy). — *Geophysical Journal International* **196/2**, 633-643.
- Caricchi, L., Petrelli, M., Bali, E., Sheldrake, T., Pioli, L. & Simpson, G. 2020: A Data Driven Approach to Investigate the Chemical Variability of Clinopyroxenes From the 2014–2015 Holuhraun–Bárdarbunga Eruption (Iceland). — *Frontiers in Earth Science* **8**, 18.
- Carniel, R. 2014: Characterization of volcanic regimes and identification of significant transitions using geophysical data: a review. — *Bulletin of Volcanology* **76/8**, 1-22.
- Carniel, R. & Guzmán, S. R. 2020: Machine Learning in Volcanology: A Review. — *Volcanoes-Updates in Volcanology*.
- Carniel, R., Di Cecca, M. & Jaquet, O. 2006: A user-friendly, dynamic web environment for remote data browsing and analysis of multiparametric geophysical data within the MULTIMO project. — *Journal of volcanology and geothermal research* **153/1-2**, 80-96.
- Carniel, R., Jolis, E. M. & Jones, J. 2010: A geophysical multi-parametric analysis of hydrothermal activity at Dallol, Ethiopia. — *Journal of African Earth Sciences* **58/5**, 812-819.
- Carniel, R., Barbui, L. & Jolly, A. 2013: Detecting dynamical regimes by Self-Organizing Map (SOM) analysis: an example from the March 2006 phreatic eruption at Raoul Island, New Zealand Kermadec Arc. — *Bollettino di Geofisica Teorica ed Applicata* **54/1**.
- Carniel, R., Cabras, G., Ichihara, M. & Takeo, M. 2014: Filtering wind in infrasound data by non-negative matrix factorization. — *Seismological Research Letters* **85/5**, 1056-1062.
- Cassisi, C., Prestifilippo, M., Cannata, A., Montalto, P., Patane, D. & Privitera, E. 2016: Probabilistic reasoning over seismic time series: Volcano monitoring by hidden Markov models at Mt. Etna. — *Pure and Applied Geophysics* **173/7**, 2365-2386.
- Charrad, M., Ghazzali, N., Boiteau, V. & Niknafs, A. 2014: Determining the best number of clusters in a data set. — *J Stat Softw.*
- Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. 2002: SMOTE: synthetic minority over-sampling technique. — *Journal of artificial intelligence research* **16**, 321-357.

- Cheng, L. & Costa, F. 2019: Statistical analysis of crystal populations and links to volcano deformation for more robust estimates of magma replenishment volumes. — *Geology* **47/12**, 1171-1175.
- Cheng, L., Costa, F. & Carniel, R. 2017: Unraveling the presence of multiple plagioclase populations and identification of representative two-dimensional sections using a statistical and numerical approach. — *American Mineralogist* **102/9**, 1894-1905.
- Chouet, B. 2003: Volcano seismology. — *Pure and applied geophysics* **160/3**, 739-788.
- Christ, M., Braun, N., Neuffer, J. & Kempa-Liehr, A. W. 2018: Time series feature extraction on basis of scalable hypothesis tests (tsfresh—a python package). — *Neurocomputing* **307**, 72-77.
- Coppola, D., Laiolo, M., Cigolini, C., Massimetti, F., Delle Donne, D., Ripepe, M., Arias, H., Barsotti, S., Parra, C. B. & Centeno, R. G. 2020: Thermal remote sensing for global volcano monitoring: Experiences from the MIROVA system. — *Frontiers in Earth Science* **7**, 362.
- Corradino, C., Ganci, G., Cappello, A., Bilotta, G., Hérault, A. & Del Negro, C. 2019: Mapping recent lava flows at Mount Etna using multispectral sentinel-2 images and machine learning techniques. — *Remote Sensing* **11/16**, 1916.
- Corradino, C., Ganci, G., Cappello, A., Bilotta, G., Calvari, S. & Del Negro, C. 2020: Recognizing eruptions of Mount Etna through machine learning using multiperspective infrared images. — *Remote Sensing* **12/6**, 970.
- Corsaro, R. A., Falsaperla, S. & Langer, H. 2013: Geochemical pattern classification of recent volcanic products from Mt. Etna, Italy, based on Kohonen maps and fuzzy clustering. — *International journal of earth sciences* **102/4**, 1151-1164.
- Cortes, C. & Vapnik, V. 1995: Support-vector networks Machine learning (pp. 237–297), Vol. 20. — *Boston, MA: Kluwer Academic Publisher*.
- Cortés, G., Carniel, R., Ángeles Mendoza, M. & Lesage, P. 2019: Standardization of noisy volcanoseismic waveforms as a key step toward station-independent, robust automatic recognition. — *Seismological Research Letters* **90/2A**, 581-590.
- Cortés, G., Benítez, M. C., García, L., Álvarez, I. & Ibanez, J. M. 2015: A comparative study of dimensionality reduction algorithms applied to volcano-seismic signals. — *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **9/1**, 253-263.
- Cortés, G., Carniel, R., Lesage, P., Mendoza, M. & Della Lucia, I. 2020: Practical Volcano-Independent Recognition of Seismic Events: VULCAN. ears Project. — *Frontiers in Earth Science* **8**, 702.
- Cortés, G., Arámbula, R., Gutiérrez, L., Benítez, C., Ibáñez, J., Lesage, P., Alvarez, I. & Garcia, L. 2009: Evaluating robustness of a HMM-based classification system of volcano-seismic events at colima and popocatepetl volcanoes. — 2009 IEEE International Geoscience and Remote Sensing Symposium. IEEE, pp. II-1012-II-1015.
- Cortés, J. A., Palma, J. L. & Wilson, M. 2007: Deciphering magma mixing: the application of cluster analysis to the mineral chemistry of crystal populations. — *Journal of Volcanology and Geothermal Research* **165/3-4**, 163-188.
- Costa, F., Shea, T. & Ubide, T. 2020: Diffusion chronometry and the timescales of magmatic processes. — *Nature Reviews Earth & Environment* **1/4**, 201-214.
- Curilem, G., Vergara, J., Fuentealba, G., Acuña, G. & Chacón, M. 2009: Classification of seismic signals at Villarrica volcano (Chile) using neural networks and genetic algorithms. — *Journal of volcanology and geothermal research* **180/1**, 1-8.
- Curilem, M., Huenupan, F., San Martin, C., Fuentealba, G., Cardona, C., Franco, L., Acuña, G. & Chacón, M. 2014a: Feature analysis for the classification of volcanic seismic events using support vector machines. — Mexican International Conference on Artificial Intelligence. Springer, pp. 160-171.
- Curilem, M., Vergara, J., San Martin, C., Fuentealba, G., Cardona, C., Huenupan, F., Chacón, M., Khan, M. S., Hussein, W. & Yoma, N. B. 2014b: Pattern recognition applied to seismic signals of the Llaima volcano (Chile): An analysis of the events' features. — *Journal of volcanology and geothermal research* **282**, 134-147.
- Curilem, M., Huenupan, F., Beltrán, D., San Martin, C., Fuentealba, G., Franco, L., Cardona, C., Acuña, G., Chacón, M. & Khan, M. S. 2016: Pattern recognition applied to seismic signals of Llaima volcano

- (Chile): An evaluation of station-dependent classifiers. — *Journal of Volcanology and Geothermal Research* **315**, 15-27.
- Currenti, G., Del Negro, C. & Nunnari, G. 2005: Inverse modelling of volcanomagnetic fields using a genetic algorithm technique. — *Geophysical Journal International* **163/1**, 403-418.
- Dabrowa, A., Green, D., Rust, A. & Phillips, J. 2011: A global study of volcanic infrasound characteristics and the potential for long-range monitoring. — *Earth and Planetary Science Letters* **310/3-4**, 369-379.
- Dawson, P. B., Benitez, M., Chouet, B. A., Wilson, D. & Okubo, P. G. 2010: Monitoring very-long-period seismicity at Kilauea Volcano, Hawaii. — *Geophysical Research Letters* **37/18**.
- Del Pezzo, E., Esposito, A., Giudicepietro, F., Marinaro, M., Martini, M. & Scarpetta, S. 2003: Discrimination of earthquakes and underwater explosions using neural networks. — *Bulletin of the Seismological Society of America* **93/1**, 215-223.
- Delicado, P. 2008: Comparing methods for dimensionality reduction when data are density functions.
- Dempsey, D., Cronin, S. J., Mei, S. & Kempa-Liehr, A. W. 2020: Automatic precursor recognition and real-time forecasting of sudden explosive volcanic eruptions at Whakaari, New Zealand. — *Nature communications* **11/1**, 1-8.
- Donnadieu, F. & Merle, O. 1998: Experiments on the indentation process during cryptodome intrusions: New insights into Mount St. Helens deformation. — *Geology* **26/1**, 79-82.
- Duque, A., González, K., Pérez, N., Benítez, D., Grijalva, F., Lara-Cueva, R. & Ruiz, M. 2020: Exploring the unsupervised classification of seismic events of Cotopaxi volcano. — *Journal of Volcanology and Geothermal Research* **403**, 107009.
- Dye, B. C. & Morra, G. 2020: Machine learning as a detection method of Strombolian eruptions in infrared images from Mount Erebus, Antarctica. — *Physics of the Earth and Planetary Interiors* **305**, 106508.
- Efremenko, D. S., Loyola R, D. G., Hedelt, P. & Spurr, R. J. 2017: Volcanic SO₂ plume height retrieval from UV sensors using a full-physics inverse learning machine algorithm. — *International Journal of Remote Sensing* **38/sup1**, 1-27.
- El Naqa, I. & Murphy, M. J. 2015: What is machine learning? — *Machine learning in radiation oncology*. Springer, pp. 3-11.
- Esposito, A., Giudicepietro, F., Scarpetta, S., D'auria, L., Marinaro, M. & Martini, M. 2006: Automatic discrimination among landslide, explosion-quake, and microtremor seismic signals at Stromboli volcano using neural networks. — *Bulletin of the Seismological Society of America* **96/4A**, 1230-1240.
- Esposito, A. M., Alaia, G., Giudicepietro, F., Pappalardo, L. & D'antonio, M. 2021: Unsupervised geochemical analysis of the eruptive products of Ischia, Vesuvius and Campi Flegrei. — *Progresses in Artificial Intelligence and Neural Systems*. Springer, pp. 175-184.
- Esposito, A. M., Giudicepietro, F., D'auria, L., Scarpetta, S., Martini, M. G., Coltelli, M. & Marinaro, M. 2008: Unsupervised neural analysis of very-long-period events at Stromboli volcano using the self-organizing maps. — *Bulletin of the Seismological Society of America* **98/5**, 2449-2459.
- Falsaperla, S., Graziani, S., Nunnari, G. & Spampinato, S. 1996: Automatic classification of volcanic earthquakes by using multi-layered neural networks. — *Natural Hazards* **13/3**, 205-228.
- Federico, C., Aiuppa, A., Favara, R., Gurrieri, S. & Valenza, M. 2004: Geochemical monitoring of groundwaters (1998–2001) at Vesuvius volcano (Italy). — *Journal of Volcanology and Geothermal Research* **133/1-4**, 81-104.
- Fukui, K., Ando, S., Fujiwara, F., Kitagawa, S., Kokubo, K., Onizawa, S. Y., Sakai, T., Shimbori, T., Takagi, A. & Yamamoto, T. 2013: MaGCAP-V: a windows-based software to analyze ground deformation and geomagnetic change in volcanic areas. — *Abstr IAVCEI* **4**, 2C-P8.
- Gaddes, M., Hooper, A. & Bagnardi, M. 2019: Using machine learning to automatically detect volcanic unrest in a time series of interferograms. — *Journal of Geophysical Research: Solid Earth* **124/11**, 12304-12322.
- Ganseccki, C., Lee, R. L., Shea, T., Lundblad, S. P., Hon, K. & Parcheta, C. 2019: The tangled tale of Kilauea's 2018 eruption as told by geochemical monitoring. — *Science* **366/6470**.
- Géron, A. 2019: Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. — *O'Reilly Media*.

- Goldberg, D. E. & Holland, J. H. 1988: Genetic algorithms and machine learning.
- Goodfellow, I., Bengio, Y., Courville, A. & Bengio, Y. 2016: Deep learning (Vol. 1, No. 2). — Cambridge: MIT press.
- Green, D. N. & Neuberg, J. 2006: Waveform classification of volcanic low-frequency earthquake swarms and its implication at Soufrière Hills Volcano, Montserrat. — *Journal of Volcanology and Geothermal Research* **153/1-2**, 51-63.
- Green, R. M., Bebbington, M. S., Cronin, S. J. & Jones, G. 2013: Geochemical precursors for eruption repose length. — *Geophysical Journal International* **193/2**, 855-873.
- Hajian, A., Cannavò, F., Greco, F. & Nunnari, G. 2019: Classification of Mt Etna (Italy) volcanic activity by machine learning approaches. — *Annals of Geophysics* **62/2**, 231.
- Hammer, C., Beyreuther, M. & Ohrnberger, M. 2012: A seismic-event spotting system for volcano fast-response systems. — *Bulletin of the Seismological Society of America* **102/3**, 948-960.
- Handelman, G. S., Kok, H. K., Chandra, R. V., Razavi, A. H., Huang, S., Brooks, M., Lee, M. J. & Asadi, H. 2019: Peering into the black box of artificial intelligence: evaluation metrics of machine learning methods. — *American Journal of Roentgenology* **212/1**, 38-43.
- Hastie, T., Tibshirani, R. & Friedman, J. 2009: *The elements of statistical learning: data mining, inference, and prediction*. — Springer Science & Business Media.
- Hellweg, M. 2000: Physical models for the source of Lascar's harmonic tremor. — *Journal of Volcanology and Geothermal Research* **101/1-2**, 183-198.
- Hibert, C., Provost, F., Malet, J.-P., Maggi, A., Stumpf, A. & Ferrazzini, V. 2017: Automatic identification of rockfalls and volcano-tectonic earthquakes at the Piton de la Fournaise volcano using a Random Forest algorithm. — *Journal of Volcanology and Geothermal Research* **340**, 130-142.
- Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P. & Sainath, T. N. 2012: Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. — *IEEE Signal processing magazine* **29/6**, 82-97.
- Hopfield, J. J. 1982: Neural networks and physical systems with emergent collective computational abilities. — *Proceedings of the national academy of sciences* **79/8**, 2554-2558.
- Hotovec-Ellis, A. & Jeffries, C. 2016: Near real-time detection, clustering, and analysis of repeating earthquakes: Application to Mount St. Helens and Redoubt volcanoes. — Seismological Society of America Annual Meeting.
- Hughes, G. 1968: On the mean accuracy of statistical pattern recognizers. — *IEEE transactions on information theory* **14/1**, 55-63.
- Ibáñez, J. M., Benítez, C., Gutiérrez, L. A., Cortés, G., García-Yeguas, A. & Alguacil, G. 2009: The classification of seismo-volcanic signals using Hidden Markov Models as applied to the Stromboli and Etna volcanoes. — *Journal of Volcanology and Geothermal Research* **187/3-4**, 218-226.
- Ibs-Von Seht, M. 2008: Detection and identification of seismic signals recorded at Krakatau volcano (Indonesia) using artificial neural networks. — *Journal of Volcanology and Geothermal Research* **176/4**, 448-456.
- Inguaggiato, S., Diliberto, I. S., Federico, C., Paonita, A. & Vita, F. 2018: Review of the evolution of geochemical monitoring, networks and methodologies applied to the volcanoes of the Aeolian Arc (Italy). — *Earth-Science Reviews* **176**, 241-276.
- Iványi, A. 2006: *Angol-magyar informatikai szótár*. — Tinta.
- Jankovics, M. É., Harangi, S., Németh, K., Kiss, B. & Ntaflos, T. 2015: A complex magmatic system beneath the Kisssomlyó monogenetic volcano (western Pannonian Basin): evidence from mineral textures, zoning and chemistry. — *Journal of Volcanology and Geothermal Research* **301**, 38-55.
- Jankovics, M. É., Sági, T., Astbury, R. L., Petrelli, M., Kiss, B., Ubide, T., Németh, K., Ntaflos, T. & Harangi, S. 2019: Olivine major and trace element compositions coupled with spinel chemistry to unravel the magmatic systems feeding monogenetic basaltic volcanoes. — *Journal of Volcanology and Geothermal Research* **369**, 203-223.

- Jankowski, S., Szymanski, Z., Currenti, G., Napoli, R., Del Negro, C. & Fortuna, L. 2010: Modeling Volcanomagnetic Dynamics by Recurrent Least-Square Support Vector Machines. — *From Physics To Control Through An Emergent View*. World Scientific, pp. 213-218.
- Janoušek, V., Farrow, C. M. & Erban, V. 2006: Interpretation of whole-rock geochemical data in igneous geochemistry: introducing Geochemical Data Toolkit (GCDkit). — *Journal of Petrology* **47/6**, 1255-1259.
- Jaquet, O. & Carniel, R. 2003: Multivariate stochastic modelling: towards forecasts of paroxysmal phases at Stromboli. — *Journal of Volcanology and Geothermal Research* **128/1-3**, 261-271.
- Jaquet, O., Carniel, R., Sparks, S., Thompson, G., Namar, R. & Di Cecca, M. 2006: DEVIN: A forecasting approach using stochastic methods applied to the Soufriere Hills Volcano. — *Journal of Volcanology and Geothermal Research* **153/1-2**, 97-111.
- Johnson, J. B., Anderson, J., Marcillo, O. & Arrowsmith, S. 2012: Probing local wind and temperature structure using infrasound from Volcan Villarrica (Chile). — *Journal of Geophysical Research: Atmospheres* **117/D17**.
- Johnston, M. 1997: Review of electric and magnetic fields accompanying seismic and volcanic activity. — *Surveys in Geophysics* **18/5**, 441-476.
- Jones, J., Carniel, R. & Malone, S. 2012: Subband decomposition and reconstruction of continuous volcanic tremor. — *Journal of volcanology and geothermal research* **213**, 98-115.
- Jordan, M. I. & Mitchell, T. M. 2015: Machine learning: Trends, perspectives, and prospects. — *Science* **349/6245**, 255-260.
- Jousset, P., Pallister, J., Boichu, M., Buongiorno, M. F., Budisantoso, A., Costa, F., Andreastuti, S., Prata, F., Schneider, D. & Clarisse, L. 2012: The 2010 explosive eruption of Java's Merapi volcano — a '100-year' event. — *Journal of volcanology and geothermal research* **241**, 121-135.
- Kahl, M., Chakraborty, S., Pompilio, M. & Costa, F. 2015: Constraints on the nature and evolution of the magma plumbing system of Mt. Etna Volcano (1991–2008) from a combined thermodynamic and kinetic modelling of the compositional record of minerals. — *Journal of Petrology* **56/10**, 2025-2068.
- Kennard, R. W. & Stone, L. A. 1969: Computer aided design of experiments. — *Technometrics* **11/1**, 137-148.
- Ketkar, N. 2017: Introduction to pytorch. — *Deep learning with python*. Springer, pp. 195-208.
- Kohonen, T. 2001: Learning vector quantization. — *Self-organizing maps*. Springer, pp. 245-261.
- Kotsarenko, A., Molchanov, O., Hayakawa, M., Koshevaya, S., Grimalsky, V., Pérez Enríquez, R. & López Cruz-Abeyro, J. 2005: Investigation of ULF magnetic anomaly during Izu earthquake swarm and Miyakejima volcano eruption at summer 2000, Japan. — *Natural Hazards and Earth System Sciences* **5/1**, 63-69.
- Kovács, J. & Eröss, A. 2017: Statistically optimal grouping using combined cluster and discriminant analysis (CCDA) on a geochemical database of thermal karst waters in Budapest. — *Applied Geochemistry* **84**, 76-86.
- Kovács, J., Kovács, S., Magyar, N., Tanos, P., Hatvani, I. G. & Anda, A. 2014: Classification into homogeneous groups using combined cluster and discriminant analysis. — *Environmental modelling & software* **57**, 52-59.
- Kovács, J., Kovács, S., Hatvani, I. G., Magyar, N., Tanos, P., Korponai, J. & Blaschke, A. P. 2015: Spatial Optimization of Monitoring Networkson the Examples of a River, a Lake-Wetland System and a Sub-Surface Water System. — *Water resources management* **29/14**, 5275-5294.
- Köhler, A., Ohrnberger, M. & Scherbaum, F. 2010: Unsupervised pattern recognition in continuous seismic wavefield records using self-organizing maps. — *Geophysical Journal International* **182/3**, 1619-1630.
- Kriesel, D. 2007: A brief introduction on neural networks.
- Kuhn, M. & Johnson, K. 2019: Feature engineering and selection: A practical approach for predictive models. — *CRC Press*.
- Lacassie, J. P., Del Solar, J. R., Roser, B. & Hervé, F. 2006: Visualization of volcanic rock geochemical data and classification with artificial neural networks. — *Mathematical Geology* **38/6**, 697-710.

- Lahr, J. C., Chouet, B. A., Stephens, C. D., Power, J. A. & Page, R. A. 1994: Earthquake classification, location, and error analysis in a volcanic environment: Implications for the magmatic system of the 1989–1990 eruptions at Redoubt Volcano, Alaska. — *Journal of Volcanology and Geothermal Research* **62/1-4**, 137-151.
- Langer, H., Falsaperla, S. & Thompson, G. 2003: Application of artificial neural networks for the classification of the seismic transients at Soufriere Hills volcano, Montserrat. — *Geophysical research letters* **30/21**.
- Langer, H., Falsaperla, S., Powell, T. & Thompson, G. 2006: Automatic classification and a-posteriori analysis of seismic event identification at Soufriere Hills volcano, Montserrat. — *Journal of volcanology and geothermal research* **153/1-2**, 1-10.
- Langer, H., Falsaperla, S., Masotti, M., Campanini, R., Spampinato, S. & Messina, A. 2009: Synopsis of supervised and unsupervised pattern classification techniques applied to volcanic tremor data at Mt Etna, Italy. — *Geophysical Journal International* **178/2**, 1132-1144.
- Lara, P. E. E., Fernandes, C. a. R., Inza, A., Mars, J. I., Métaixian, J.-P., Dalla Mura, M. & Malfante, M. 2020: Automatic multichannel volcano-seismic classification using machine learning and EMD. — *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **13**, 1322-1331.
- Lecocq, T., Caudron, C. & Brenguier, F. 2014: MSNoise, a python package for monitoring seismic velocity changes using ambient seismic noise. — *Seismological Research Letters* **85/3**, 715-726.
- Lever, J., Krzywinski, M. & Altman, N. 2016: Regularization. — Nature Publishing Group.
- López, M. Á., Caricchi, L., Jorgenson, C., Musu, A. & Giordano, G. 2021: The transitions from mildly explosive to caldera-forming eruptions at Colli Albani volcano (Italy). — *Copernicus Meetings*.
- Lukács, R., Guillong, M., Bachmann, O., Fodor, L. & Harangi, S. 2021a: Tephrostratigraphy and magma evolution based on combined zircon trace element and U-Pb age data: Fingerprinting Miocene silicic pyroclastic rocks in the Pannonian basin. — *Frontiers in Earth Science* **9**, 64.
- Lukács, R., Petrelli, M., Guillong, M., Bachmann, O., Fodor, L. & Harangi, S. 2021b: *The zircon proxy in tephrostratigraphy and magma evolution studies. Fingerprinting Miocene silicic pyroclastic rocks in the Pannonian Basin and its surroundings*. — Copernicus Meetings.
- Lunardon, N., Menardi, G. & Torelli, N. 2014: ROSE: A Package for Binary Imbalanced Learning. — *R journal* **6/1**.
- Malfante, M., Dalla Mura, M., Métaixian, J.-P., Mars, J. I., Macedo, O. & Inza, A. 2018: Machine learning for volcano-seismic signals: Challenges and perspectives. — *IEEE Signal Processing Magazine* **35/2**, 20-30.
- Manley, G. F., Pyle, D. M., Mather, T. A., Rodgers, M., Clifton, D. A., Stokell, B. G., Thompson, G., Londoño, J. M. & Roman, D. C. 2020: Understanding the timing of eruption end using a machine learning approach to classification of seismic time series. — *Journal of Volcanology and Geothermal Research* **401**, 106917.
- Marchetti, E. & Ripepe, M. 2005: Stability of the seismic source during effusive and explosive activity at Stromboli Volcano. — *Geophysical research letters* **32/3**.
- Marchetti, E., Ripepe, M., Campus, P., Le Pichon, A., Vergoz, J., Lacanna, G., Mialle, P., Héreil, P. & Husson, P. 2019: Long range infrasound monitoring of Etna volcano. — *Scientific reports* **9/1**, 1-10.
- Marcillo, O. & Johnson, J. 2010: Tracking near-surface atmospheric conditions using an infrasound network. — *The Journal of the Acoustical Society of America* **128/1**, EL14-EL19.
- Masotti, M., Falsaperla, S., Langer, H., Spampinato, S. & Campanini, R. 2006: Application of Support Vector Machine to the classification of volcanic tremor at Etna, Italy. — *Geophysical research letters* **33/20**.
- Masotti, M., Campanini, R., Mazzacurati, L., Falsaperla, S., Langer, H. & Spampinato, S. 2008: TREMOReC: a software utility for automatic classification of volcanic tremor. — *Geochemistry, Geophysics, Geosystems* **9/4**.
- Massimetti, F., Diego, C., Marco, L., Sébastien, V., Cigolini, C. & Ripepe, M. 2020: Volcanic Hot-Spot Detection Using SENTINEL-2: A Comparison with MODIS-MIROVA Thermal Data Series. — *Remote Sensing* **12/5**, 820.

- Matoza, R., Fee, D., Green, D. & Mialle, P. 2019: Volcano infrasound and the international monitoring system. — *Infrasound monitoring for atmospheric studies*. Springer, pp. 1023-1077.
- Mccullagh, P. & Nelder, J. 1989: Generalized Linear Models Second edition Chapman & Hall. — London.
- McNutt, S. R. 2002: Volcano seismology and monitoring for eruptions. — *International Geophysics Series* **81/A**, 383-406.
- Minsky, M. & Papert, S. 1969: Perceptron: an introduction to computational geometry. — *The MIT Press, Cambridge, expanded edition* **19/88**, 2.
- Mitchell, T. M. 2006: *The discipline of machine learning*. — 9. Carnegie Mellon University, School of Computer Science, Machine Learning
- Molina, I., Kumagai, H. & Yepes, H. 2004: Resonances of a volcanic conduit triggered by repetitive injections of an ash-laden gas. — *Geophysical Research Letters* **31/3**.
- Murphy, K. P. 2012: *Machine learning: a probabilistic perspective*. — MIT press.
- Murwantara, I. M., Yugopuspito, P. & Hermawan, R. 2020: Comparison of machine learning performance for earthquake prediction in Indonesia using 30 years historical data. — *Telkomnika* **18/3**, 1331-1342.
- Musu, A., Caricchi, L., Perugini, D., Corsaro, R. A., Vetere, F. & Petrelli, M. 2021: Crystal zoning patterns: competition between crystal growth and elements diffusion. — *Copernicus Meetings*.
- Nunnari, G. & Del Negro, C. 2003: Inverse Modelling of magnetic data by using a neural network approach. —: EGS-AGU-EUG Joint Assembly, pp. 9591.
- Ohrnberger, M. 2001: Continuous automatic classification of seismic signals of volcanic origin at Mt. Merapi, Java, Indonesia. — *Dissertation*, 168.
- Palacio-Niño, J.-O. & Berzal, F. 2019: Evaluation metrics for unsupervised learning algorithms. — *arXiv preprint arXiv:1905.05667*.
- Pan, S. J. & Yang, Q. 2009: A survey on transfer learning. — *IEEE Transactions on knowledge and data engineering* **22/10**, 1345-1359.
- Park, I., Jolly, A., Lokmer, I. & Kennedy, B. 2020: Classification of long-term very long period (VLP) volcanic earthquakes at Whakaari/White Island volcano, New Zealand. — *Earth, Planets and Space* **72/1**, 1-10.
- Pearce, J. A. & Cann, J. R. 1973: Tectonic setting of basic volcanic rocks determined using trace element analyses. — *Earth and planetary science letters* **19/2**, 290-300.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R. & Dubourg, V. 2011: Scikit-learn: Machine learning in Python. — *the Journal of machine Learning research* **12**, 2825-2830.
- Petrelli, M., Caricchi, L. & Perugini, D. 2020: Machine Learning Thermo-Barometry: Application to Clinopyroxene-Bearing Magmas. — *Journal of Geophysical Research: Solid Earth* **125/9**, e2020JB020130.
- Petrelli, M., Bizzarri, R., Morgavi, D., Baldanza, A. & Perugini, D. 2017: Combining machine learning techniques, microanalyses and large geochemical datasets for tephrochronological studies in complex volcanic areas: New age constraints for the Pleistocene magmatism of central Italy. — *Quaternary Geochronology* **40**, 33-44.
- Pham, B. T., Phong, T. V., Nguyen-Thoi, T., Parial, K., K. Singh, S., Ly, H.-B., Nguyen, K. T., Ho, L. S., Le, H. V. & Prakash, I. 2020: Ensemble modeling of landslide susceptibility using random subspace learner and different decision tree classifiers. — *Geocarto International*, 1-23.
- Pignatelli, A. & Piochi, M. 2021: Machine learning applied to rock geochemistry for predictive outcomes: The Neapolitan volcanic history case. — *Journal of Volcanology and Geothermal Research*, 107254.
- Poland, M. P. & Anderson, K. R. 2020: Partly cloudy with a chance of lava flows: Forecasting volcanic eruptions in the twenty-first century. — *Journal of Geophysical Research: Solid Earth* **125/1**, e2018JB016974.
- Rabiner, L. R. 1989: A tutorial on hidden Markov models and selected applications in speech recognition. — *Proceedings of the IEEE* **77/2**, 257-286.

- Ramis, R. O., Garcia, A., Marrero, J. M., De La Cruz-Reyna, S., Carniel, R. & Vila, J. 2019: Volcanic and volcano-tectonic activity forecasting: a review on seismic approaches. — *Annals of Geophysics* **62/1**, 06.
- Rasband, W. S. 1997: ImageJ. — Bethesda, MD.
- Reitermanova, Z. 2010: Data splitting. — WDS, pp. 31-36.
- Ren, C. X., Peltier, A., Ferrazzini, V., Rouet-Leduc, B., Johnson, P. & Brenguier, F. 2020: Machine learning reveals the seismic signature of eruptive behavior at Piton de la Fournaise volcano. — *Geophysical research letters* **47/3**, e2019GL085523.
- Rosenblatt, F. 1958: The perceptron: a probabilistic model for information storage and organization in the brain. — *Psychological review* **65/6**, 386.
- Rouwet, D. & Tassi, F. 2011: Geochemical monitoring of volcanic lakes. A generalized box model for active crater lakes. — *Annals of Geophysics* **54/2**.
- Royston, P. & White, I. R. 2011: Multiple imputation by chained equations (MICE): implementation in Stata. — *J Stat Softw* **45/4**, 1-20.
- Rumelhart, D. E., Hinton, G. E. & Williams, R. J. 1986: Learning representations by back-propagating errors. — *nature* **323/6088**, 533-536.
- Russell Stuart, J. & Norvig, P. 2009: *Artificial intelligence: a modern approach*. — Prentice Hall.
- Sarbas, B. 2008: The GEOROC database as part of a growing geoinformatics network. — *Geoinformatics 2008—Data to Knowledge*. USGS, pp. 42-43.
- Scarpetta, S., Giudicepietro, F., Ezin, E. C., Petrosino, S., Del Pezzo, E., Martini, M. & Marinaro, M. 2005: Automatic classification of seismic signals at Mt. Vesuvius volcano, Italy, using neural networks. — *Bulletin of the Seismological Society of America* **95/1**, 185-196.
- Sevilla, W. I., Jumawan, L. A., Clarito, C. J., Quintia, M.-A., Dominguiano, A. A. & Solidum Jr, R. U. 2020: Improved 1D velocity model and deep long-period earthquakes in Kanlaon Volcano, Philippines: Implications for its magmatic system. — *Journal of Volcanology and Geothermal Research* **393**, 106793.
- Shabtai, A., Elovici, Y. & Rokach, L. 2012: A survey of data leakage detection and prevention solutions. — *Springer Science & Business Media*.
- Shanmugamani, R. 2018: Deep Learning for Computer Vision: Expert techniques to train advanced neural networks using TensorFlow and Keras. — *Packt Publishing Ltd*.
- Si, A., Zhang, J., Zhang, Y., Kazuva, E., Dong, Z., Bao, Y. & Rong, G. 2020: Debris flow susceptibility assessment using the integrated random forest based steady-state infinite slope method: A case study in Changbai Mountain, China. — *Water* **12/7**, 2057.
- Sivia, D. & Skilling, J. 2006: Data analysis: a Bayesian tutorial. — OUP Oxford.
- Snee, R. D. 1977: Validation of regression models: methods and examples. — *Technometrics* **19/4**, 415-428.
- Snow, C. A. 2006: A reevaluation of tectonic discrimination diagrams and a new probabilistic approach using large geochemical databases: Moving beyond binary and ternary plots. — *Journal of Geophysical Research: Solid Earth* **111/B6**.
- Sutton, R. S. & Barto, A. G. 2018: Reinforcement learning: An introduction. — MIT press.
- Tan P., Steinbach M. & Kumar V. 2011: Bevezetés az adatbányászatba — Panem Kft.
- Tárraga, M., Carniel, R., Ortiz, R. & García, A. 2008: The failure forecast method: Review and application for the real-time detection of precursory patterns at reawakening volcanoes. — *Developments in Volcanology* **10**, 447-469.
- Tharwat, A. 2016: Principal component analysis-a tutorial. — *International Journal of Applied Pattern Recognition* **3/3**, 197-240.
- Tharwat, A., Gaber, T., Ibrahim, A. & Hassanien, A. E. 2017: Linear discriminant analysis: A detailed tutorial. — *AI communications* **30/2**, 169-190.

- Titos, M., Bueno, A., García, L. & Benítez, C. 2018a: A deep neural networks approach to automatic recognition systems for volcano-seismic events. — *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **11/5**, 1533-1544.
- Titos, M., Bueno, A., García, L., Benítez, M. C. & Ibañez, J. 2018b: Detection and classification of continuous volcano-seismic signals with recurrent neural networks. — *IEEE Transactions on Geoscience and Remote Sensing* **57/4**, 1936-1948.
- Trnkoczy, A. 2009: Understanding and parameter setting of STA/LTA trigger algorithm. — *New Manual of Seismological Observatory Practice (NMSOP)*. Deutsches GeoForschungsZentrum GFZ, pp. 1-20.
- Trujillo-Castrillón, N., Valdés-González, C. M., Arámbula-Mendoza, R. & Santacoloma-Salguero, C. C. 2018: Initial processing of volcanic seismic signals using Hidden Markov Models: Nevado del huila, colombia. — *Journal of Volcanology and Geothermal Research* **364**, 107-120.
- Ueki, K., Hino, H. & Kuwatani, T. 2018: Geochemical discrimination and characteristics of magmatic tectonic settings: A machine-learning-based approach. — *Geochemistry, Geophysics, Geosystems* **19/4**, 1327-1347.
- Uyeda, S., Hayakawa, M., Nagao, T., Molchanov, O., Hattori, K., Orihara, Y., Gotoh, K., Akinaga, Y. & Tanaka, H. 2002: Electric and magnetic phenomena observed before the volcano-seismic activity in 2000 in the Izu Island Region, Japan. — *Proceedings of the National Academy of Sciences* **99/11**, 7352-7355.
- Valade, S., Ley, A., Massimetti, F., D'hondt, O., Laiolo, M., Coppola, D., Loibl, D., Hellwich, O. & Walter, T. R. 2019: Towards global volcano monitoring using multisensor sentinel missions and artificial intelligence: The MOUNTS monitoring system. — *Remote Sensing* **11/13**, 1528.
- Vapnik, V. 1998: The support vector method of function estimation. — *Nonlinear modeling*. Springer, pp. 55-85.
- Verma, S. P. & Díaz-González, L. 2020: New discriminant-function-based multidimensional discrimination of mid-ocean ridge and oceanic plateau. — *Geoscience Frontiers* **11/5**, 1681-1693.
- Voight, B. & Elsworth, D. 2000: Instability and collapse of hazardous gas-pressurized lava domes. — *Geophysical Research Letters* **27/1**, 1-4.
- Walker, J., Lehnert, K., Hofmann, A., Sarbas, B. & Carlson, R. 2005: EarthChem: International collaboration for solid earth geochemistry in geoinformatics. — AGU Fall Meeting Abstracts, pp. IN44A-03.
- Wassermann, J., Krischer, L., Megies, T., Barsch, R. & Beyreuther, M. 2013: ObsPy: A Python toolbox for seismology. — AGU Fall Meeting Abstracts, pp. S51A-2322.
- Watson, L. M. 2020: Using unsupervised machine learning to identify changes in eruptive behavior at Mount Etna, Italy. — *Journal of Volcanology and Geothermal Research* **405**, 107042.
- Widrow, B. 1960: Adaptive "adaline" Neuron Using Chemical "memistors".
- Williams, M. J., Schoneveld, L., Mao, Y., Klump, J., Gosses, J., Dalton, H., Bath, A. & Barnes, S. 2020: pyrolite: Python for geochemistry. — *Journal of Open Source Software* **5/50**, 2314.
- Witsil, A. J. & Johnson, J. B. 2020a: Analyzing continuous infrasound from Stromboli volcano, Italy using unsupervised machine learning. — *Computers & Geosciences* **140**, 104494.
- Witsil, A. J. & Johnson, J. B. 2020b: Volcano video data characterized and classified using computer vision and machine learning algorithms. — *Geoscience Frontiers* **11/5**, 1789-1803.
- Wolpert, D. H. 1996: The lack of a priori distinctions between learning algorithms. — *Neural computation* **8/7**, 1341-1390.
- Wujek, B., Hall, P. & Günes, F. 2016: Best practices for machine learning applications. — *SAS Institute Inc.*
- Yang, G., Ireland, D., Kaiser, R. & Mahon, D. 2018: Machine Learning for Muon Imaging. — International Conference on Brain Inspired Cognitive Systems. Springer, pp. 808-817.
- Zhang, S., Zhang, C. & Yang, Q. 2003: Data preparation for data mining. — *Applied artificial intelligence* **17/5-6**, 375-381.
- Zhou, Y.-T. & Chellappa, R. 1988: Computation of optical flow using a neural network. — *ICNN*, pp. 71-78.

Zhu, X. & Goldberg, A. B. 2009: Introduction to semi-supervised learning. — *Synthesis lectures on artificial intelligence and machine learning* **3/1**, 1-130.

Zlotnicki, J., Bof, M., Perdereau, L., Yvetot, P., Tjetjep, W., Sukhyar, R. & Purbawinata, M. 2000: Magnetic monitoring at Merapi volcano, Indonesia. — *Journal of volcanology and geothermal research* **100/1-4**, 321-336.

Zobin, V. M. 2012: Introduction to volcanic seismology. — 6. Elsevier.

Szakdolgozat:

Miklós, M. M., József, K., & István, M. (2018). Gépi tanulási algoritmusok használata kisalföldi sekélyfúrási adatok automatikus osztályozására.— MSc szakdolgozat. ELTE TTK, Földrajz Földtudományi Intézet

Internetes források:

Brownlee, J. (2019). How Much Training Data is Required for Machine Learning. Machine Learning Mastery. In: <https://machinelearningmastery.com/much-training-data-required-machine-learning/> (Letöltés: 2021. 03. 09.)

Brownlee, J. (2020). Parametric and Nonparametric Machine Learning Algorithms. Machine Learning Mastery. In: <https://machinelearningmastery.com/parametric-and-nonparametric-machine-learning-algorithms/> (Letöltés: 2021. 03. 20.)

Glander, S. (2017). Dealing with Unbalanced Data in Machine Learning. R-Bloggers. In: <https://www.r-bloggers.com/2017/04/dealing-with-unbalanced-data-in-machine-learning/> (Letöltés: 2021. 03. 09.)

Jordan, J. (2020). Effective testing for machine learning systems. Data Science. In: <https://www.jeremyjordan.me/testing-ml/> (Letöltés: 2021. 04. 23.)

Jordan, J. (2017). Evaluating a machine learning model. Data Science. In: <https://www.jeremyjordan.me/evaluating-a-machine-learning-model/> (Letöltés: 2021. 03. 09.)

Mahendru, K. (2019). A Detailed Guide to 7 Loss Functions for Machine Learning Algorithms with Python Code. Analytics Vidhya. In: <https://www.analyticsvidhya.com/blog/2019/08/detailed-guide-7-loss-functions-machine-learning-python-code/> (Letöltés: 2021. 03. 09.)

Voskoglou, C. (2017). What is the best programming language for Machine Learning. Towards Data Science, 5. In: <https://towardsdatascience.com/what-is-the-best-programming-language-for-machine-learning-a745c156d6b7> (Letöltés: 2021. 03. 09.)

Training, Testing and Evaluating Machine Learning Models. Qualcomm Developer Network. In: <https://developer.qualcomm.com/software/qualcomm-neural-processing-sdk/learning-resources/training-testing-machine-learning-models/training-testing-evaluating-machine-learning-models> (Letöltés: 2021. 03. 09.)

MIROVA projekt (Firenzei Egyetem és Torinói Egyetem együttműködése): <http://www.mirovaweb.it/> (Letöltés: 2021. 04. 28.)

VULCAN.ears, EU Horizon 2020, 749249. számú projekt: <https://cordis.europa.eu/project/id/749249> (Letöltés: 2021. 04. 28.)