

Analiza zbioru danych "Patient-level dataset to study the effect of COVID-19 in people with Multiple Sclerosis" z wykorzystaniem nadzorowanego algorytmu uczenia maszynowego One-Class SVM(Support Vector Machines)

Anna Mrozek, Bartosz Panek i Przemysław Jura

ARTICLE INFO

Keywords:
One-Class SVM
COVID-19

STRESZCZENIE

Artykuł analizuje zbiór danych pacjentów ze stwardnieniem rozsianym (SM), którzy przeszli COVID-19, przy użyciu nadzorowanego algorytmu One-Class SVM. Celem jest identyfikacja przypadków o zwiększonym ryzyku ciężkiego przebiegu infekcji. Umożliwi to zidentyfikowanie charakterystycznych cech pacjentów bardziej narażonych na powikłania.

1. Wprowadzenie

Pandemia COVID-19 miała znaczący wpływ na osoby z chorobami przewlekłymi, w tym na pacjentów ze stwardnieniem rozsianym (SM). SM jest przewlekłą chorobą autoimmunologiczną, która wpływa na układ nerwowy i może prowadzić do trwałego uszkodzenia neuronów oraz ograniczenia funkcji motorycznych oraz poznawczych.[1][2] Ze względu na charakter choroby i stosowane terapie immunosupresyjne, pacjenci z SM mogą być bardziej narażeni na cięższy przebieg infekcji wirusowych, w tym COVID-19.[3][4] Analiza danych od pacjentów z SM, którzy przeszli COVID-19, może dostarczyć ważnych informacji na temat ryzyka powikłań i zidentyfikować czynniki przyczyniające się do poważniejszych objawów, co ostatecznie może pomóc w lepszym monitorowaniu i opiece nad tą grupą pacjentów.

2. Cel

Celem badania jest wykorzystanie nadzorowanego algorytmu One-Class SVM do analizy zbioru danych pacjentów ze stwardnieniem rozsianym, którzy przeszli COVID-19, w celu zidentyfikowania cech pacjentów o zwiększonym ryzyku ciężkiej infekcji. Oczekuje się, że wyniki analizy dostarczą wskazówek do opracowania bardziej ukierunkowanych strategii opieki i środków zapobiegawczych dla pacjentów ze stwardnieniem rozsianym w kontekście potencjalnych przyszłych pandemii lub innych zagrożeń wirusowych.

3. Przegląd literatury

W literaturze medycznej i naukowej, od początku pandemii COVID-19, wiele badań skupiało się na analizie wpływu wirusa SARS-CoV-2 na osoby cierpiące na choroby przewlekłe i autoimmunologiczne, takie jak stwardnienie rozsiane (SM). Wyniki tych badań sugerują, że pacjenci z SM, zwłaszcza ci stosujący terapie immunosupresyjne,

mogą być bardziej narażeni na cięższy przebieg COVID-19 oraz związane z nim powikłania[5][6]. U pacjentów z SM często obserwuje się obniżoną odporność oraz większą podatność na infekcje, co wynika zarówno z samej choroby, jak i z efektów leków tłumiących układ odpornościowy[7]. Dodatkowo w badaniach zwraca się uwagę na takie czynniki ryzyka jak wiek, płeć, rodzaj stosowanego leczenia oraz inne choroby towarzyszące, które mogą zwiększać ryzyko ciężkiego przebiegu infekcji u tej grupy pacjentów.

W ostatnich latach coraz większą popularność zyskuje zastosowanie algorytmów uczenia maszynowego, takich jak Support Vector Machines (SVM), do analizy danych medycznych. Algorytmy te pozwalają na wykrywanie wzorców i anomalii w dużych, zróżnicowanych zbiorach danych[8]. W szczególności algorytm One-Class SVM, używany do wykrywania anomalii, jest przydatny do identyfikacji przypadków wysokiego ryzyka w danych medycznych, gdzie przypadki nietypowe (np. pacjenci bardziej narażeni na ciężki przebieg COVID-19) występują rzadko. Badania pokazują, że One-Class SVM dobrze sprawdza się w analizie populacji pacjentów, gdy dostęp do dużych zbiorów danych o przypadkach zdrowych jest ograniczony, co często stanowi wyzwanie w analizie medycznej[9].

W kontekście badań nad COVID-19 i SM, algorytmy nadzorowanego uczenia maszynowego, takie jak SVM, umożliwiają dokładniejszą analizę czynników ryzyka oraz lepsze prognozowanie ciężkiego przebiegu choroby. Wyniki takich analiz mogą być przydatne nie tylko dla lekarzy, ale także dla decydentów zajmujących się zdrowiem publicznym, ponieważ pozwalają na opracowanie bardziej ukierunkowanych strategii opieki dla osób z SM, które są bardziej narażone na zagrożenia związane z wirusami, takimi jak COVID-19.

4. Metodologia

W badaniu zastosowano algorytm nadzorowanego uczenia maszynowego One-Class SVM w celu identyfikacji pacjentów z podwyższonym ryzykiem ciężkiego przebiegu COVID-19 w grupie osób ze stwardnieniem rozsianym.

Analiza obejmowa a cechy kliniczne pacjent w, takie jak wiek, p c, typ leczenia oraz choroby w sp istniej ce, aby zidentyfikowa  wzorce zwi zane z wi ksz  podatno ci  na powik ania.

4.1. Dataset

Stwardnienie rozsiane (MS) to przewlek a choroba autoimmunologiczna, kt ra wywo uje stan zapalny w obr bie o rodkowego uk adu nerwowego. Choroba prowadzi do r żnych stopni utraty funkcji przez uszkodzenia mieliny oraz w  kien nerwowych [10]. Osoby z MS s  bardziej podatne na infekcje z powodu z o zonego dzia ania samej choroby, jej leczenia i naturalnego przebiegu [11]. Z inicjatywy COVID-19 and MS Global Data Sharing Initiative (GDSI) zbadano, jak leki immunosupresyjne lub immunomoduluj ce wp ywaj  na COVID-19 i jego przebieg u os b z MS. GDSI mia a na celu zwi kszenie skali zbierania danych dotycz cych COVID-19 i dostarczenie spo eczno ci zwi zanej z MS informacji opartych na danych podczas pandemii [12]. W ramach GDSI wybrano kluczowe zmienne obejmuj ce informacje o COVID-19, stopniu jego ci  ko ci, leczeniu, dane demograficzne, historię i nasilenie MS, stosowanie lekw modyfikuj cych przebieg choroby (DMT), choroby w sp istniej ce i wybrane zachowania zwi zane ze stylem  ycia, takie jak palenie tytoniu. Globalna spo eczno c MS w sp pracowa a, przekazuj c dokumentacj  statusu COVID-19 u os b z MS za po rednictwem centralnej platformy udost pnionej przez QMENTA [13].

Ten zbi r danych zosta  zebrany za pomoc  nar dzia do szybkiego wprowadzania danych, kt re umo liwia o klinicystom, osobom ze stwardnieniem rozsianym (PwMS) lub ich przedstawicielom bezpo rednie wprowadzanie informacji do centralnej platformy GDSI. Nar dzie to zawiera o kwestionariusz oparty na wcze niej ustalonych zmiennych i nie gromadzi o bezpo rednich danych osobowych, aby chroni c prywatno c u ytkownik w. Nar dzie zosta  wy  czone 3 lutego 2022 roku.

Zbi r danych obejmuje informacje o 1141 osobach ze stwardnieniem rozsianym (PwMS). Aby zapewni c zgodno c danych z wytycznymi HIPAA, przeprowadzono proces deidentyfikacji. Po zebraniu danych dokonano oceny ryzyka zwi zanego z ma ymi kom rkami (SCRA), klasyfikuj c zmienne na trzy kategorie: bezpo rednie identyfikatory, zmienne wra liwe i identyfikatory po rednie. Bezpo rednie identyfikatory to zmienne, kt re mog  jednoznacznie zidentyfikowa  osob , zmienne wra liwe to te, kt re respondent mo e chcie  zachowa  w tajemnicy, natomiast identyfikatory po rednie mog  zidentyfikowa  osob , je li s  po  czone z danymi z innych zbior w.

Poniewa  w danych nie zbierano imion pacjent w, deidentyfikacja skupia a si  na datach i wieku pacjent w. Daty w kolumnie „stop-or-end-date-combined” zosta  przesuni te o losow  liczb  dni (mi dzy -15 a 15), aby uniemo liwi c identyfikacj  na podstawie dat. Wiek pacjent w sklasyfikowano w cztery grupy: 0 dla os b w wieku 0–17 lat, 1 dla os b mi dzy 18 a 50 lat, 2 dla os b mi dzy 51 a 70 lat, oraz 3 dla os b powy żej 71 lat. Dzi ki temu  adne dok adne

warto ci wieku powy żej 90 lat nie zosta  ujawnione. Po klasyfikacji zmiennych i wdro eniu odpowiednich  rodk w ostro no ci dane zosta  zdeidentyfikowane i spe niaj  standardy HIPAA, zachowuj c jednocze nie warto c badawcz . Ponadto, aby zapewni c ochron  prywatno ci, zastosowano techniki takie jak K-anonimizacja oraz r żnorodno c.

Zbi r danych obejmuje zestaw z g ry okre lonych zmiennych ($n=47$), takich jak p c, kategoria wiekowa, typ MS, wynik EDSS, status palenia oraz kategoria BMI. Te zmienne dostarczaj  informacji o demografii pacjent w, ich stanie klinicznym oraz symptomach zwi zanych z COVID-19. Szczeg owy opis typ w zmiennych i ich statystyki znajduje si  w sekcji „Opis Danych”.

4.2. Opis metody

Jednoklasowy SVM (One-Class Support Vector Machine, OCSVM) to algorytm przeznaczony do wykrywania anomalii w zbiorze danych. Zamiast klasyfikowa  dane do dw ch lub wi cej klas, jak w klasycznym SVM, OCSVM koncentruje si  wy cznie na danych normalnych.[14][15] Jego celem jest zbudowanie granicy wok l wi kszo ci przypadkw normalnych, tworz c „stref  normalno ci”, kt ra odr bnia standardowe przypadki od anomalii.

G wna zasada dzia ania OCSVM opiera si  na maksymalizacji marginesu mi dzy przypadkami normalnymi a granic , kt ra oddziela norm  od anomalii. Dzi ki temu model lepiej rozpoznaje dane odstaj ce, kt re znajduj  si  poza wyznaczon  stref . W algorytmie znajduje si  hiperparametr „nu”, kt ry pozwala kontrolowa  czu o c modelu – decyduje on o maksymalnym odsetku b d w marginesowych oraz liczbie wektor w no nych, wp ywaj c na balans mi dzy surowo ci  a tolerancj  modelu.[16][17]

Kluczowym elementem dzia ania OCSVM jest funkcja decyzyjna oparta na hiperp aszczy nie, kt ra oddziela dane od pocz tku uk adu w sp r dnych. Wz r na hiperp aszczy n  wyrażany jest jako:

$$f(x) = \mathbf{w} \cdot \mathbf{x} - \rho \quad (1)$$

gdzie:

- \mathbf{w} – wektor normalny do hiperp aszczy zny, wyznaczony podczas procesu uczenia,
- \mathbf{x} – pr bka danych,
- ρ – jest warto ci  progow  (bias), kt ra definiuje margines.

OCSVM posiada szereg parametr w, dzi ki kt rym mo liwe jest dostosowywanie modelu do specyficznych potrzeb:

- **Parametr ν :** Kontroluje liczb  obserwacji uznawanych za anomalie. Parametr ten przyjmuje warto ci w przedziale od 0 do 1 i determinuje udzia  anomalii, jakie model mo e tolerowa  w zbiorze treningowym (np. je li $\nu = 0.05$, oznacza to,  e oko o 5% pr bek zostanie sklasyfikowanych jako anomalie).

- **Parametr γ :** Wpływa na kształt granicy decyzyjnej poprzez dopasowywanie modelu do danych. Mniejsza wartość γ oznacza „szerszą” granicę, obejmującą więcej punktów, natomiast większa wartość γ powoduje, że granica jest bardziej dopasowana do danych, co może zwiększać ryzyko nadmiernego dopasowania (overfitting).
- **Jądro (*ang. kernel*):** One-Class SVM często wykorzystuje funkcje jądrowe, aby modelować nieliniowe granice decyzyjne i lepiej dopasować się do danych, które nie są liniowo rozdzielne w przestrzeni cech. Wybór jądra, zwany również „sztuczką jądra”, wpływa na charakter dopasowania modelu i jego wydajność.

Podczas treningu OCSVM analizuje wyłącznie dane normalne, co sprawia, że jest szczególnie użyteczny w sytuacjach, gdy anomalie są rzadkie lub trudne do zidentyfikowania. Model staje się wtedy bardziej niezawodny w rzeczywistych zastosowaniach, takich jak wykrywanie oszustw, monitorowanie awarii czy zabezpieczanie sieci komputerowych. OCSVM może wykorzystać różne funkcje jądra, co umożliwia mu wykrywanie zarówno prostych, jak i bardziej złożonych odchyłań.[18]

4.3. Opis przeprowadzonych obliczeń

1. Przygotowanie i czyszczenie danych - rozpoczęliśmy od wczytania danych oraz wybrania interesujących nas cech, które następnie zostały przekształcone w sposób umożliwiający przeprowadzenie obliczeń:

- Dla kolumn binarnych (yes/no) wartości zostały zamienione na liczby 0 i 1.
- Dla kolumn kategorycznych i ordinalnych (np. age-in-cat, covid19-outcome-levels-2, report-source) zostały przypisane wartości liczbowe, zgodnie z przyjętym mapowaniem.
- Brakujące wartości w danych zostały wypełnione zerami, co pozwoliło na uniknięcie problemów podczas analizy i treningu modelu.

2. Tworzenie nowych zmiennych: symptom-score oraz comorbidity-score, aby skonsolidować informacje o objawach i chorobach współistniejących.

- Stworzyliśmy dwie kolumny symptom-score oraz comorbidity-score, które zliczają ilość objawów wirusa COVID-19 oraz liczby chorób współistniejących dla każdego pacjenta.
- Kolumny te były następnie skalowane przy użyciu StandardScaler, co pozwala na lepszą interpretację wyników oraz standaryzację w zakresie modelowania.

Efektem było, uzyskanie wartości numerycznych, które reprezentują intensywność symptomów oraz liczbę chorób współistniejących dla każdego pacjenta.

3. Przygotowanie danych do treningu - wykorzystaliśmy pacjentów bez chorób współistniejących jako grupę uczącą (zbiór X-train). Wszystkie próbki (zarówno te z chorobami współistniejącymi, jak i bez) tworzyły zestaw testowy (zbiór X-test). Wybór pacjentów bez chorób współistniejących pozwolił modelowi nauczyć się cech charakterystycznych dla „normalnych” próbek, które posłużyły do wykrywania potencjalnych anomalii.

4. Trening modelu One-Class SVM - zastosowaliśmy algorytm One-Class Support Vector Machine (SVM) z jądrem rbf, który dobrze radzi sobie z modelowaniem nieliniowych granic decyzyjnych. Model trenowano, aby rozpoznawał typowe cechy pacjentów bez chorób współistniejących, a następnie przewidywał, które próbki są podobne do tego wzorca, a które mogą być anomaliami. Po treningu model przypisywał każdej próbce etykietę 1 (normalna) lub -1 (anomia). Próbki oznaczone jako anomalia zawarto w nowej ramce anomalies w celu dalszej analizy.

5. Analiza wyników - na podstawie wykrytych anomalii przeprowadziliśmy szereg analiz, tworząc wizualizacje, które pomagają zrozumieć charakterystykę anomalii[19]

4.4. Wykorzystane metryki oceny

Classification Report: Zawiera dokładność (precision), czułość (recall) i f1-score dla obu klas (1 - normalna, -1 - anomalia). Daje to pogląd na efektywność wykrywania anomalii oraz odsetek poprawnie sklasyfikowanych normalnych próbek.

Confusion Matrix (Macierz Pomyłek): Macierz błędów wyświetla liczby klasyfikacji poprawnych i błędnych dla klas normalnych i anomalii, co pozwala zidentyfikować potencjalne problemy z fałszywie pozytywnymi lub fałszywie negatywnymi klasyfikacjami.

ROC AUC (Area Under the Curve of Receiver Operating Characteristic) — pomimo braku klasy binarnej, można skonstruować ROC, badając, jak dobrze model oddziela „normę” od „odchyłań”.

Procent wykrytych anomalii — określa, jaki odsetek faktycznych anomalii został prawidłowo wykryty.

False Positive Rate (FPR) — aby zminimalizować fałszywe alarmy.

5. Wyniki

6. Opis wyników

7. Podsumowanie

8. Bibliografia

- [1] <https://diag.pl/pacjent/artykuly/jakie-sa-pierwsze-objawy-i-sposoby-leczenia-stwardnienia-rozsianego/>
- [2] <https://www.medicover.pl/o-zdrowiu/stwardnienie-rozsiane-objawy-przyczyny-i-leczenie,6619,n,192>
- [3] https://pl.wikipedia.org/wiki/Stwardnienie_rozsiane
- [4] <https://ptsr.org.pl/strona/133,covid-19-a-sm>
- [5] Sormani MP, De Rossi N, Schiavetti I, Carmisciano L, Cordioli C, Radaelli M, et al. Disease modifying therapies

Ocena modelu One-Class SVM:

	precision	recall	f1-score	support
Normal	1.00	1.00	1.00	37
Anomaly	1.00	1.00	1.00	47
accuracy			1.00	84
macro avg	1.00	1.00	1.00	84
weighted avg	1.00	1.00	1.00	84

Macierz pomyłek:

```
[[37  0]
 [ 0 47]]
```

ROC AUC: 1.0

Procent wykrytych anomalii (TPR): 100.00%

False Positive Rate (FPR): 0.00%

- [18] <https://medium.com/@roshmitadey/anomaly-detection-using-support-vectors-2c1b842213ed>
- [19] <https://github.com/przemyslawJura00/RozpoznawanieWzorcow>

Figure 1: podpis

and COVID-19 severity in multiple sclerosis. *Ann Neurol.* 2021 Apr

[6] Louapre C, Collongues N, Stankoff B, Giannesini C, Papeix C, Bensa C, et al. Clinical characteristics and outcomes in patients with coronavirus disease 2019 and multiple sclerosis. *JAMA Neurol.* 2020 Sep

[7] Simpson-Yap S, De Brouwer E, Kalincik T, Rijke N, Hillert JA, Walton C, et al. Associations of Disease-Modifying Therapies With COVID-19 Severity in Multiple Sclerosis. *Neurology.* 2021 Nov

[8] Schiff MA, Rae-Grant A, Gilden D, Franklin GM. Practice guideline: Disease-modifying therapies for adults with multiple sclerosis: Report of the guideline development, dissemination, and implementation subcommittee of the American Academy of Neurology. *Neurology.* 2019 Jan

[9] Erfani P, Mitchell AJ, Hameed S, Heydarpour P, Ghaffaripour R, Sahraian MA. Systematic review of health-related quality of life in multiple sclerosis patients: The impact of pharmacological treatments and lifestyle. *J Neurol Sci.* 2016 Dec

[10] Calabresi PA. Diagnosis and management of multiple sclerosis. *Am Fam Physician.* 2004 Nov

[11] Montgomery S, Hillert J, Bahmanyar S. Hospital admission due to infections in multiple sclerosis patients. *Eur J Neurol.* 2013 Aug

[12] Peeters LM, Parciak T, Walton C, Geys L, Moreau Y, De Brouwer E, et al. COVID-19 in people with multiple sclerosis: A global data sharing initiative. *Mult Scler Houndmills Basingstoke Engl.* 2020 Sep

[13] Simpson-Yap S, De Brouwer E, Kalincik T, Rijke N, Hillert JA, Walton C, et al. Associations of Disease-Modifying Therapies With COVID-19 Severity in Multiple Sclerosis. *Neurology.* 2021 Nov

[14] <https://physionet.org/content/patient-level-data-covid-ms/1.0.1/>

[15] <https://www.geeksforgeeks.org/understanding-one-class-support-vector-machines/>

[16] <https://scikit-learn.org/dev/modules/generated/sklearn.svm.OneClassSVM.html>

[17] <https://www.baeldung.com/cs/one-class-svm>