

## **DECODE - zastosowanie analizy statystycznej oraz metod maszynowego uczenia w celu opisu pacjentów chorych na COVID-19 oraz predykcji prawdopodobieństwa wystąpienia choroby COVID-19 na bazie prostych cech klinicznych**

**Aleksandra Gruca<sup>1</sup>, Małgorzata Bach<sup>2</sup>, Paweł Foszner<sup>3</sup>, Joanna Henzel<sup>1</sup>, Mateusz Kania<sup>4</sup>, Michał Kozielski<sup>1</sup>, Justyna Mika<sup>4</sup>, Anna Papież<sup>4</sup>, Joanna Tobiasz<sup>4</sup>, Aleksandra Werner<sup>2</sup>, Joanna Żyła<sup>4</sup>, Jerzy Jaroszewicz<sup>5</sup>, Joanna Polańska<sup>4</sup>, Marek Sikora<sup>1,6</sup>**

<sup>1</sup> Katedra Sieci i Systemów Komputerowych, Wydział Automatyki, Elektroniki i Informatyki, Politechnika Śląska, Gliwice

<sup>2</sup> Katedra Informatyki Stosowanej, Wydział Automatyki, Elektroniki i Informatyki, Politechnika Śląska, Gliwice

<sup>3</sup> Katedra Grafiki, Wizji Komputerowej i Systemów Cyfrowych, Wydział Automatyki, Elektroniki i Informatyki, Politechnika Śląska, Gliwice

<sup>4</sup> Katedra Inżynierii i Analizy Eksploracyjnej Danych, Wydział Automatyki, Elektroniki i Informatyki, Politechnika Śląska, Gliwice

<sup>5</sup> Katedra i Oddział Kliniczny Chorób Zakaźnych i Hepatologii, Śląski Uniwersytet Medyczny, Bytom

<sup>6</sup> Sieć Badawcza Łukasiewicz – Instytut Technik Innowacyjnych EMAG, Katowice

Celem projektu DECODE – Data drivEn COVID DEtection jest zbudowanie bazy danych pacjentów z podejrzeniem COVID-19, a następnie o wykorzystanie metod analizy statystycznej oraz metod maszynowego uczenia w celu stworzenia modeli zarówno eksploracyjnych jak i predykcyjnych pozwalających na powiązanie cech klinicznych z diagnozą COVID-19.

We współpracy z Oddziałem Obserwacyjno-Zakaźnym i Hepatologii Szpitala Specjalistycznego nr 1 w Bytomiu zebrano dane dotyczące ponad 2 tysięcy pacjentów, którzy w okresie od marca do czerwca 2020 roku zgłosili się na oddział obserwacyjno-zakaźny szpitala z podejrzeniem zakażenia chorobą COVID-19. Dla każdego z pacjentów zebrano podstawowe informacje demograficzne, informacje dotyczące szeregu możliwych objawów choroby COVID-19, chorób współistniejących oraz przyjmowanych leków. Każdy z pacjentów przebadany został testem RT-PCR na obecność wirusa SARS-CoV-2. Na podstawie objawów klinicznych oraz wyników testu pacjenci podzielni zostali na grupy SARS-CoV-2(+) z objawami chorobowymi, SARS-CoV-2(+) bez objawów chorobowych, SARS-CoV-2(-) z objawami chorobowymi, SARS-CoV-2(-) bez objawów chorobowych.

Pierwszym celem projektu DECODE jest wykorzystanie zarówno narzędzi analizy statystycznej jak i metod maszynowego uczenia w celu identyfikacji cech charakteryzujących poszczególne grupy pacjentów. Analizowane są tu zarówno dane demograficzne, objawy kliniczne jak i choroby współistniejące. Stworzenie modeli wiążących cechy pacjentów z wynikiem testu pozwoli na zrozumienie różnic pomiędzy analizowanymi grupami, a w szczególności odpowiedź na pytanie czy da się stworzyć charakterystykę pacjenta chorego oraz pacjenta zdrowego.

Kolejnym celem jest stworzenie modelu predykcyjnego, który na podstawie występowania objawów mógłby dostarczać informacji o prawdopodobieństwie, iż dany pacjent jest chory na COVID-19. Model taki pozwoliłby w sposób przesiewowy identyfikować pacjentów potencjalnie chorych i wymagających dalszego testowania. Z uwagi na konieczność wychwycenia prawie wszystkich przypadków osób chorych na COVID-19, najważniejszą cechą tworzonych modeli predykcyjnych będzie ich wysoka czułość oraz wysoka wartość wskaźnika NPV.

Poza oczywistą próbą stworzenia modeli odróżniających pacjentów będących nosicielami wirusa SARS-CoV-2 od osób z wynikiem negatywnym testu na SARS-CoV-2 projekt DECODE usiłuje również

odpowiedzieć na pytanie czy istnieją jakieś cechy charakteryzujące bezobjawowych nosicieli wirusa SARS-CoV-2 w porównaniu do osób zakażonych mających objawy infekcji.