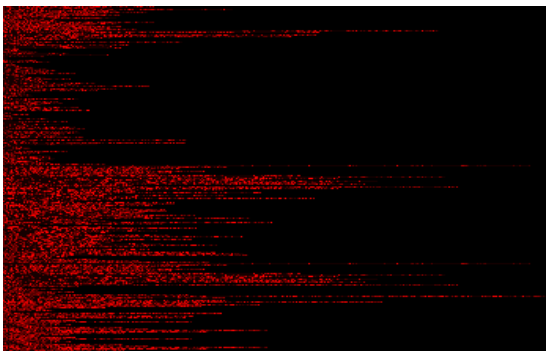
Katedra Systemów Multimedialnych

Zespół projektowy: 30@KSMM'2021	**1. Franciszek Górski - kierownik 2. Piotr Juszczyk 3. Sebastian Lewalski**
Opiekun:	**mgr inż. Sebastian Cygert**
Klient:	**dr hab. inż. Piotr Szczuko (Konsorcjum AI Tech - Akademia innowacyjnych zastosowań technologii cyfrowych)**
Data zakończenia:	**01.2022**
Słowa kluczowe:	**Tumor-educated platelets, Cancer Detection, Liquid Biopsy, DNA-Seq, Convolutional Neural Network, Decision Tree**



TEMAT PROJEKTU:

Zastosowanie algorytmów uczenia maszynowego do wykrywania raka w płynnej biopsji

CELE I ZAKRES PROJEKTU:

Celem projektu jest wykorzystanie algorytmów uczenia maszynowego do wykrywania raka na podstawie danych pobranych z płynnej biopsji.

W zakres prac wchodzą:

1. Wykorzystanie algorytmów opartych o sieci neuronowe, drzewa decyzyjne lub inne techniki uczenia maszynowego do klasyfikacji danych z płynnej biopsji.
2. Analiza i redukcja zbioru danych na podstawie istotności cech.
3. Zastosowanie wcześniej wykorzystanych algorytmów na zredukowanym zbiorze danych i porównanie wyników.

OSIĄGNIĘTE REZULTATY:

1. Wyniki uzyskane przez architekturę ResNet w wersji 18 i 34 warstwowej dają rezultaty powyżej 90% metryki balanced accuracy dla zbioru walidacyjnego i testowego. Są to wyniki zadowalające, zbliżone do porównywanych przez nas wyników zespołu z GUMedu.
2. Wyniki uzyskane z wykorzystaniem algorytmów drzew decyzyjnych dają rezultaty powyżej 80% metryki balanced accuracy dla zbioru walidacyjnego i testowego. Wyniki w dużym stopniu zależą od wartości parametru określającego maksymalną głębokość pozostałych drzew decyzyjnych.
3. Redukcja zbioru danych na podstawie istotności cech.
4. Redukcja danych w oparciu o istotność cech dała wyraźną poprawę w osiąganych rezultatach przez sieci CNN. Poprawa wyniosła kilkanaście punktów procentowych, co sugeruje, że redukcja danych może być niezbędna w procesie wstępnego przetwarzania.

CECHY CHARAKTERYSTYCZNE ROZWIĄZANIA, KIERUNKI DALSZYCH PRAC:

Cechy charakterystyczne:

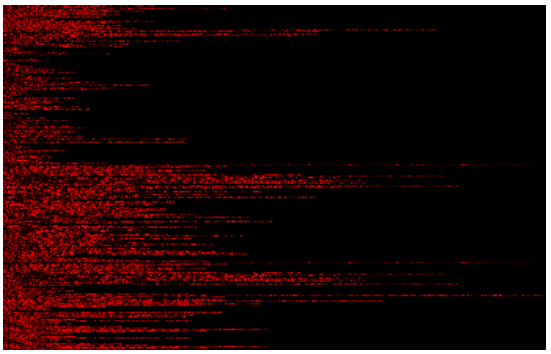
1. Zbiór danych jest mocno niezbalansowany, co powoduje, że do pomiaru skuteczności modelu sieci neuronowych wykorzystano odpowiednie metryki takie jak recall, specificity czy balanced accuracy.
2. Modele powstałe przy użyciu algorytmów drzew decyzyjnych dobrze różnicują próbki zdrowe od chorych i grupy kontrolnej. Niewielkie głębokości drzew decyzyjnych (rzędu 1-2) lepiej radzą sobie z klasyfikacją.
3. Odfiltrowanie szumu stanowiącego aż 75% oryginalnych danych.

Kierunki dalszych prac:

1. Optymalizacja hiperparametrów i architektur sieci konwolucyjnych.
2. Testowanie innych algorytmów drzew decyzyjnych.
3. Zastosowanie większej liczby klas w algorytmach drzew decyzyjnych.
4. Zastosowanie innych metod augmentacji danych.

DEPARTMENT FULL NAME

Project team: 30@KSMM'2021	**1. Franciszek Górski - leader 2. Piotr Juszczyk 3. Sebastian Lewalski**
Supervisor:	**mgr eng. Sebastian Cygert**
Client:	**D. Sc. eng. Piotr Szczuko (AI Tech Consortium)**
Date:	**01.2022**
Key words:	**Tumor-educated platelets, Cancer Detection, Liquid Biopsy, DNA-Seq, Convolutional Neural Network, Decision Tree**



PROJECT TITLE:

Application of machine learning algorithms for cancer detection in liquid biopsy

OBJECTIVES AND SCOPE:

The goal of the project is application of machine learning algorithms for cancer detection in data gathered from liquid biopsy.

The scope of work includes:
1. Application of algorithms based on neural networks, decision trees or other machine learning techniques for cancer detection in data gathered from liquid biopsy.
2. Analysis and reduction of the dataset based on feature relevance.
3. Application of aforementioned algorithms on a reduced dataset and comparison of the results.

RESULTS:

1. The results obtained by the ResNet architecture in the 18 and 34 layer versions yield above 90% balanced accuracy metric for the validation and test set. These are satisfactory results, similar to the GUMed team's.
2. The results of decision tree algorithms give results of above 80% balanced accuracy for the validation and test set. The results mostly depend on the value of the parameter defining the maximum depth of the decision trees.
3. Reduction of the dataset based on feature relevance.
4. Data reduction based on feature relevance yielded a significant improvement in CNNs' performance. The improvement was several percentage points, suggesting that data reduction may be necessary in preprocessing.

MAIN FEATURES, FUTURE WORKS:

1. The dataset is highly unbalanced, resulting in the use of relevant metrics such as recall, specificity, and balanced accuracy to measure the performance of the neural network model.
2. Models created with the use of decision tree algorithms distinguish healthy samples from patients and the control group well. Decision trees with small depth (1 to 2) are better at classification.
3. Filtering of noise that constituted as much as 75% of the original data.

Directions for future work:

1. Optimization of hyperparameters and convolutional network architectures.
2. Testing of other decision tree algorithms.
3. Use of more classes in decision tree algorithms
4. Application of other data augmentation methods