Ben-Gurion University of the Negev
Industrial Engineering and Management

אוניברסיטת בן-גוריון בנגב
המחלקה להנדסת תעשיה וניהול

<table>
<tr><th colspan="2">Project No.</th><th colspan="4">Project Title</th></tr>
<tr><td colspan="2">2021-01-132</td><td colspan="4">Early Diagnosis of Pulmonary Embolism Using Machine Learning</td></tr>
<tr><th colspan="3">Academic Advisor</th><th colspan="3">Co-Advisor</th></tr>
<tr><td colspan="3">Prof. Boaz Lerner</td><td colspan="3"></td></tr>
<tr><th colspan="6">Team Members</th></tr>
<tr><td colspan="2">Ori Ben Yehuda</td><td colspan="2"></td><td colspan="2"></td></tr>
<tr><td colspan="2">oribenye@post.bgu.ac.il</td><td colspan="2"></td><td colspan="2"></td></tr>
</table>

Abstract

Pulmonary embolism (PE) is a blockage of the main artery of the lung or one of its branches. Experts from the Sheba Tel HaShomer Medical Center have argued before us that many cases of PE are not diagnosed or diagnosed too late in hospitalization, when the condition becomes critical and very difficult to treat. Hence, there is a need to create a reliable data-driven model to early identify PE for a new patient.

In this study, we focus on building a machine learning (ML) based model for predicting PE at the moment a patient arrives at the hospital. We obtained from our colleagues in Sheba two data sets, one with 1,942 PE patients, and one with 44,697 patients who have never had PE. The data sets include medical records of the patients as soon as, and only when, they arrive at the emergency room, e.g., their demographics, prior diagnoses, and chronic medications taken. After pre-processing the data, and to deal with the imbalance in the data (a ratio of 1 to 23 between positive and negative to PE), we implemented a random forest (RF) model using two methods. The first method optimized the decision threshold, which determines the probability above which a patient will be classified as positive for PE using a validation set, at 0.038 (where we expected it to reflect the ratio between the priors of the two classes). The second method modeled using 28 classifiers, each is trained on a balanced set of (the same) patients and a random (with no replacement) sample of the control patients, before averaging performance over the ensemble.

Comparing the models geometric mean (GM) measure, which is the square root of the product of the true positive rate (TPR) and true negative rate (TNR), the two methods returned the same result of 0.756. Method 1 has a small advantage in TPR over Method 2 (0.667>0.656), and Method 2 has an advantage in TNR over Method 1 (0.871>0.857). As for the importance the RF model calculated for the features using the Gini index, a past PE diagnosis was ranked highest by both methods (normalized importance of 0.549 and 0.4, respectively), and then were ranked Age (0.134), BMI (0.108), and whether the patient had a previous lung disease (0.071) by the first method, and Age (0.152), whether the patient had a previous lung disease (0.145), and BMI (0.121) by the second method.

In conclusion, this work demonstrated PE prediction at hospital admission using ML models. It will be extended to further prediction during hospitalization using data collected after admission.

Keywords: Pulmonary Embolism, Machine Learning, Imbalanced data, Feature importance, Hospital admission model