Uncovering Potential Proteomic Biomarkers for Cancer Patients with COVID-19 Infection using Multilabel Deep Learning Model

Marcelo Benedeti Palermo[1], **Cristiano André da Costa**[1], **Rodrigo da Rosa Righi**[1]

[1]SOFTWARELAB, Programa de Pós-Graduação em Computacao Aplicada,
Universidade do Vale do Rio dos Sinos
Av. Unisinos 950, São Leopoldo, 93022-750, RS, Brazil

mbpalermo@edu.unisinos.br, {cac, rrr}@unisinos.br

Abstract. *The effects of COVID-19 on cancer patients are concerning. This work proposes a framework that employs a multilabel classifier processing longitudinal proteomics patients' data to identify potential proteomic biomarkers that correlate cancer and COVID-19. The framework uses Olink NPX data from 305 COVID-19-positive cancer patients. Stratified k-fold cross-validation addresses data imbalance. The overall average results show a Jaccard index of 88.79%, a hamming loss of 0.32%, a Wasserstein distance of 0.64%, and an area under the curve of 94.47%, across 312 labels, with four proteins presenting a Jaccard index of 97% or above, identified as proeminent biomarkers.*

1. Introduction

The COVID-19 disease has raised significant concerns about its implications for cancer patients. Due to immunosuppression from the disease and its treatments, cancer patients face a higher risk of severe COVID-19 outcomes [Zhou et al. 2023]. Studies suggest that COVID-19 can significantly impact this vulnerable population, increasing their risk of severe outcomes [Fung and Babik 2021]. SARS-CoV-2 infection leads to chronic inflammation, which may influence tumor behavior and prognosis. Therefore, evaluating the impact of COVID-19 on inflammatory proteins and identifying biomarkers of systemic inflammation in recovering cancer patients is essential [Kocsmár et al. 2024].

Proteomics, through blood plasma analysis, offers valuable insights into the effects of COVID-19 on cancer patients, early diagnosis, and customized treatments. Identifying proteomic biomarkers linked to COVID-19 can improve early detection and treatment strategies [Liew et al. 2024]. AI-driven analysis of proteomics accelerates the processing of large biological datasets. Machine Learning (ML) techniques help identify patterns in proteomic data, revealing biomarkers tied to COVID-19 and its long-term effects [Lv et al. 2024]. However, limited studies have explored the correlation between COVID-19 and cancer.

The present article proposes an AI-integrated longitudinal proteomics framework to investigate the COVID-19 effects on cancer patients. By using a longitudinal proteomics study design and deep learning methodology, this article contributes to AI-driven proteomics in oncology by:

- Presenting a deep-learning model that predicts COVID-19 potential biomarkers through proteome data and multilabel classification.

- Advancing AI and proteomics integration to improve cancer patient outcomes under COVID-19 influence.

The article is structured as follows: Section 2 provides an overview of related works and highlights the key differences between those approaches and our proposal. Section 3 describes the data used and the methods employed in constructing our architecture. Section 4 presents the results of our experiments, while Section 5 offers a discussion and comparison of these results with state-of-the-art approaches. Finally, Section 6 draws conclusions based on our findings.

.

2. Related Work

The effects of COVID-19 on cancer patients have raised the need for a deeper analysis of the extent of damage to the immune status. This section presents related works involving proteomics and machine learning to identify proteins correlating to cancer and COVID-19.

Hossain et al. [Hossain et al. 2024] employed traditional ML techniques to examine the impact of smoking and COVID-19 on lung cancer. They identified 10 proteins from the intersection of lung cancer (LC) and smoking and between LC and COVID-19. They tested 76 shared proteins and 10 hub proteins. Yadalam et al. [Yadalam et al. 2025] applied ML modeling to uncover novel serum proteomic biomarkers for oral squamous cell carcinoma influenced by COVID-19. The study suggested that 28 proteins showed significant differential abundance in COVID-19 patients with oral cancer compared to the control. Patel et al. [Patel et al. 2023] used targeted proteomics to compare the expression of 2925 unique blood proteins in long-COVID outpatients versus COVID-19 inpatients and healthy control subjects. They used the Boruta algorithm, which is based on Random Forest classifiers, to reduce the number of biomarkers and discard the ones obtained by chance.

While the studies referenced showcase the effectiveness of machine learning in correlating proteomics with cancer and COVID-19, they rely on targeted proteomics, which fails to capture the temporal dynamics of protein expression. In contrast, the present work utilizes a longitudinal proteomics study design to investigate the dynamic changes in protein expression over time. Additionally, it employs a deep learning-based multilabel classification approach to predict potential biomarkers, taking into account a diverse array of proteins as input, thus allowing for the analysis of large-scale proteomic data.

3. Methodology

The proposed methodology employs a framework to study a cohort of COVID-19-positive patients. The input acquires a Normalized Protein Expression (NPX) proteomic assay from the mass-spectrometry analysis, and the output returns the potential proteomic biomarkers [Wik et al. 2021]. The study cohort is from the Massachusetts General Hospital Emergency Department COVID-19 dataset, comprising 391 patients, with 305 individuals from the Olink oncology panel [Filbin et al. 2021]. The participants are all COVID-19-positive and have pre-existing symptoms, diseases (PESD), or immune conditions (IC)

such as cancer, chemotherapy, transplant, immunosuppressant use, or asplenia, according to Table 1. The PESD information is pertinent to cancer studies [Liu et al. 2023]. The proteomic NPX assay comprehends a maximum period of seven days.

Table 1. Olink Oncology Panel from MGH emergency department cohort

Variables	Category	Specifics
Cohort	305	COVID-19-positive
Unique proteins	1472	Protein expression per patient
Immune condition (IC)	0 1	Competent Compromised
Pre-existing symptoms or diseases (PESD)	1 2 3 4	Lung Kidney Respiratory symptom Fever symptom
Proteomic research blood draws	D_0 D_3 D_7 D_E	Day 0 Day 3 Day 7 Interruption

After acquiring the NPX assay from the Olink Explore HT tool, the methodology reads the NPX assay, performs the proteomic analysis to determine significant proteins (SP) involved, then uses a multi-hot encoding technique to transform the SP into binary labels. The methodology concatenates IC and PESD as features and fills gaps for any missing IC and PESD combination. The Deep Learning (DL) multilabel classifier identifies SPs as potential biomarkers (PB) for further biological investigation. This article outlines the identified PB, their corresponding UniProt identifiers, and their associated tissue specificity, providing a concise description of each biomarker [Consortium 2019]. Figure 1 outlines the framework details.

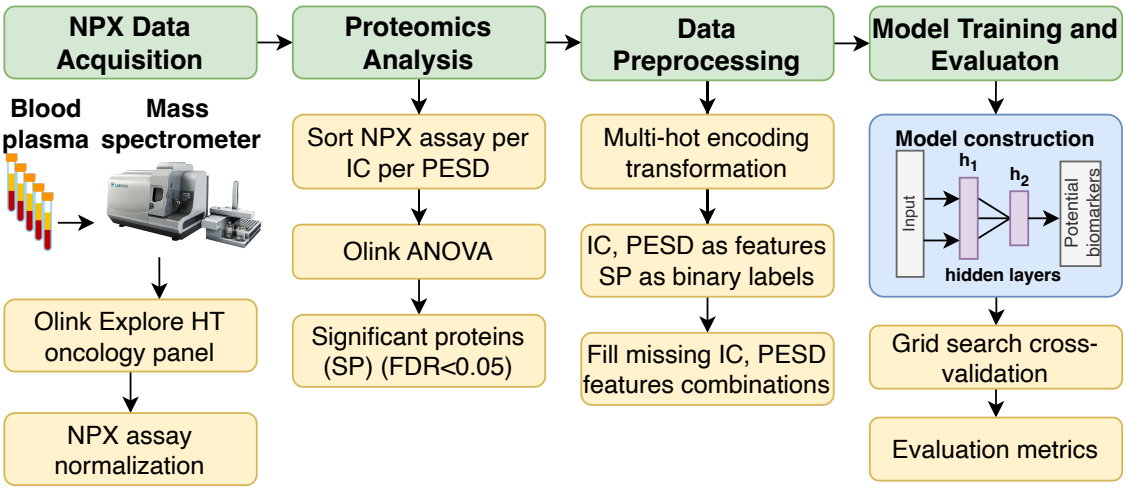

Figure 1. Framework workflow and stages proposed in this methodology.

3.1. Deep Learning Multilabel Classifier

A supervised Deep Learning (DL) multilabel classifier is the core of the workflow, predicting PB based on significant proteomic patterns in the patient cohort. The model utilizes a dataset from the Olink Analysis of Variance (ANOVA) F-test, computed using the OlinkAnalyze R package [Boberg et al. 2023], which captures the expression variances of significant proteins as input features. The DL model processes an input dataset of paired observations $\{(\mathbf{X_i}, \mathbf{Y_i})\}_{i=1}^{M}$, where $\mathbf{X}$ represents the feature set and $\mathbf{Y}$ denotes the corresponding labels. The neural network architecture consists of two hidden layers, h_1 and h_2, structured such that $|h_1| > |h_2|$. We selected a two-layer architecture to achieve a higher F1-macro score under a multilabel classification framework with the adopted cross-validation strategy.

The output layer produces predictions $\{(\hat{\mathbf{Y}}_\mathbf{i})\}_{i=1}^{N}$, where $\hat{\mathbf{Y}}$ contains the predicted labels representing the PB. The number of predicted labels matches the true labels, maintaining $N = |\mathbf{Y_i}| = |\hat{\mathbf{Y}}_\mathbf{i}|$. Let n_1 and n_2 be number of neurons in hidden layers h_1 and h_2, respectively, with $n_1 > n_2$.

3.2. Proteomic Data Analysis

The first stage involves analyzing the NPX assay data collected from the patient cohort on days D_0, D_3, D_7, and D_E. The first stage involves analyzing the NPX assay data collected from the patient cohort on days D_0, D_3, D_7, and D_E. The NPX dataset contains each patient's proteomic levels from the research blood draw. We sort the NPX assay into two arrays based on IC: one array for samples where IC is present and another for samples where IC is absent for each PESD. This sorting step helps define the features needed for the subsequent DL model training.

The stage calculates the ANOVA F-test on each sorted NPX assay to assess the variance in protein expression. The ANOVA calculation adjusts the p-values for protein expressions by applying the Benjamini-Hochberg false discovery rate (FDR) to control for multiple testing, in line with standard practices for protein significance [Kluger and Owen 2024]. Proteins with a threshold of FDR$<$0.05 are considered significant. This threshold identifies the SP that will serve as labels in the DL preprocessing stage.

3.3. Data Preprocessing

Since we use a multilabel classifier, we need to transform SP into labels because combinations of IC and PESD may point to various proteins simultaneously. We use the multi-hot encoding (MHE) strategy to achieve this task, which is present in the *MultilabelBinarizer* class from Python's sci-kit-learn package. To transform the SP names into labels, we only consider the protein names as labels because all proteins are significant. Next, the Multi-Hot Encoding (MHE) multilabel binarization technique is applied to the pre-candidate proteins identified through ANOVA, converting them into matrix binary columns [Li et al. 2022]. This MHE transformation assigns each protein a label, forming the $\mathbf{Y_i}$ label set for the multilabel classification approach.

After generating the MHE labels, the process incorporates the combinations of IC and PESD variables as feature columns, producing the ML training input dataset

$\{(\mathbf{X_i}, \mathbf{Y_i})\}_{i=1}^{M}$ with both features and labels. The final dataset forms a matrix, combining the features and MHE labels, as shown in Equation (1).

$$\{(\mathbf{X_i}, \mathbf{Y_i})\} = \begin{bmatrix} a_1 & p_1 & c_1 & s_1 & \dots & s_N \\ a_2 & p_2 & c_2 & s_1 & \dots & s_N \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ a_M & p_M & c_M & s_1 & \dots & s_N \end{bmatrix} \quad (1)$$

where M is the number of data points, N is the number of labels with $N = |\mathbf{Y_i}|$, $a \in [0, 1]$ is the IC, p represents each PESD, $c \in [0, 1]$ is the presence or absence of PESD, and $s_1, s_2, \dots, s_{N-1}, s_N$ are the labels. The final dimension of $\{(\mathbf{X_i}, \mathbf{Y_i})\}$ is $M \times (F + N)$, where F is the number of features (a, p and c).

3.4. Model Training and Evaluation

The final stage of the workflow involves training and evaluating the DL model to predict PB. Given the limited patient cohort size, this experiment employs stratified k-fold (SKF) cross-validation to mitigate bias and variance [Szeghalmy and Fazekas 2023]. The F1-macro score is favored over accuracy for multilabel classification approach, as it treats each label with equal weight [García-Pedrajas et al. 2024].

The stage processes $\{(\mathbf{X_i}, \mathbf{Y_i})\}_{i=1}^{M}$ and tests different k-folds to improve the F1-macro score. Grid search cross-validation (GSCV) fine-tunes the hyperparameters and tests sensitivity to outliers. SKF also applies balanced class weights to equalize loss penalties and reduce data imbalance. An early stopping mechanism halts the training epoch to prevent overfitting when the F1-macro score stops improving [Bai et al. 2021].

This experiment evaluates the model using the Jaccard index, Wasserstein distance (WD), hamming loss (HM), and the area under the curve-receiver operating characteristic (AUCROC), as these metrics have been shown to outperform confusion matrices in multilabel classification [Doknic and Möller 2025]. The Jaccard index measures the similarity between predicted and true labels, as per Equation (2). Labels with higher Jaccard index values denote PB. A high threshold of $J \geq$0.97 is intentionally set to identify the most relevant PB. The Wasserstein distance quantifies differences between probability distributions, where a lower value indicates higher similarity, as per Equation (3), with $d(\mathbf{Y_i}, \hat{\mathbf{Y}}_\mathbf{i})$ measuring the distance between elements $\mathbf{Y_i}$ and $\hat{\mathbf{Y}}_\mathbf{i}$, and $p(\mathbf{Y_i})$ representing the associated probability or mass for the element $\mathbf{Y_i}$.

$$J(\mathbf{Y}, \hat{\mathbf{Y}}) = \frac{1}{M} \sum_{i=1}^{M} \frac{|\mathbf{Y_i} \cap \hat{\mathbf{Y}}_\mathbf{i}|}{|\hat{\mathbf{Y}}_\mathbf{i}|} \quad (2) \qquad W(\mathbf{Y}, \hat{\mathbf{Y}}) = \sum_{i=1}^{M} d(\mathbf{Y_i}, \hat{\mathbf{Y}}_\mathbf{i}) \cdot p(\mathbf{Y_i}) \quad (3)$$

The hamming loss metric evaluates classification errors [Esti Anggraini et al. 2023]. Lower values indicate better performance, as shown in Equation (4). N is the number of labels. We also calculate the AUCROC for each label and then average the results through Equation (5), where $\mathbf{I}(\hat{\mathbf{Y}}_\mathbf{i} > \hat{\mathbf{Y}}_\mathbf{j})$ checks if the predicted value for the i-th data point is greater than that for the j-th when the true label of $\hat{\mathbf{Y}}_\mathbf{i}$ is 1 (positive) and $\hat{\mathbf{Y}}_\mathbf{j}$ is 0 (negative). M is the total number of data points.

$$H(\mathbf{Y}, \hat{\mathbf{Y}}) = \frac{1}{M} \sum_{i=1}^{M} \frac{|\mathbf{Y_i} \Delta \hat{\mathbf{Y}}_\mathbf{i}|}{|N|} \quad (4) \qquad AUC(\mathbf{Y}, \hat{\mathbf{Y}}) = \frac{1}{M} \sum_{i=1}^{M} \mathbf{I}(\hat{\mathbf{Y}}_\mathbf{i} > \hat{\mathbf{Y}}_\mathbf{j}) \quad (5)$$

The final stage's result is to identify PB based on the best evaluation metrics found during the experiment.

3.5. Architecture

To summarize the methodology, Figure 2 presents the whole architecture in the context of this article. The bioinformatics cloud platforms (BCP) like AWS and DNANexus can typically host the Repository (REP) and Microservice Consumption modules (MCM). This experiment is the MCM. We execute the experiment outside the BCP due to current environment constraints.

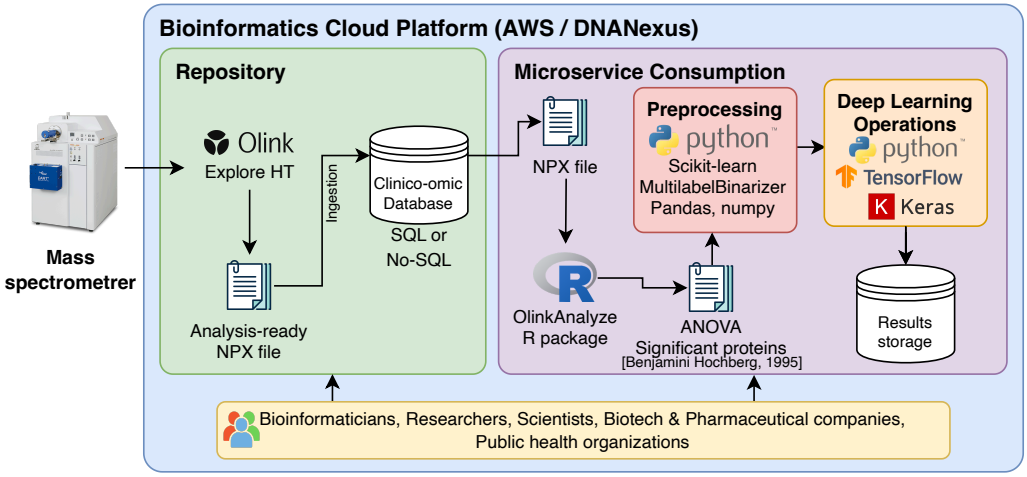

Figure 2. Proposed architecture for testing the methodology.

The mass spectrometer reads the blood plasma for the longitudinal proteomics analysis and produces the spectrum output data in raw format. The Olink Explorer HT tool normalizes the data to NPX units in the REP. The NPX normalized file is then ingested into a clinical-omic database and becomes available for the users described at the bottom of Figure 2.

The MCM reads the NPX file from the data source and applies the OlinkAnalyze R package to identify the SPs through the ANOVA F-test with FDR$<$0.05. The preprocessing utilizes the *MultilabelBinarizer* library from Python's scikit-learn package and transforms the SPs into binary labels. The *Pandas* and *numpy* libraries perform all data manipulation to concatenate the features to the labels. The resultant data serves as input to the MC. Finally, we use Tensorflow and Keras libraries to fine-tune the hyperparameters via GSCV, train the DL multilabel classifier, and issue the evaluation metrics. The identified PB are then available in a results storage.

With the proposal of the MCM architecture, we aim to contribute to early cancer diagnostic capabilities alongside COVID-19 infection management for improved patient outcomes and accelerate the understanding of the relationship between COVID-19 and cancer. Using longitudinal proteomics and deep learning, it is also possible to customize the treatment and improve the well-being of the patients infected with SARS-CoV-2 during immunotherapy.

4. Results

This section presents the results of each stage involved in the experiment based on the provided cohort.

4.1. Proteomic Data Preparation

The exploratory analysis of the cohort results in the distribution of the patients across different ICs, according to Table 2. The detailed cohort distribution per PESD and IC is available in Figure 3.

Table 2. COVID-positive cohort

Cohort	Patients
MGH dataset	391
(-) Control	8
(-) COVID-19-negative	78
Immunocompetent	214
Immunocompromised	91
Total in oncology panel	**305**

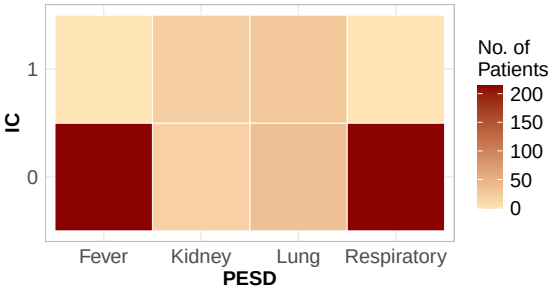

Figure 3. Patients distribution per immune status

4.2. Proteomic Results

The ANOVA F-test returns 312 SP with for FDR<0.05, as shown in Table 3 on the NPX expression assays for days D_0, D_3, D_7, and D_E. Also, the heatmap from Figure 4 shows the characteristic of the SP distribution per PESD and IC.

Table 3. Calculated significant proteins

Proteins	Totals
Significant	312
Non-significant	1160
Total	**1472**

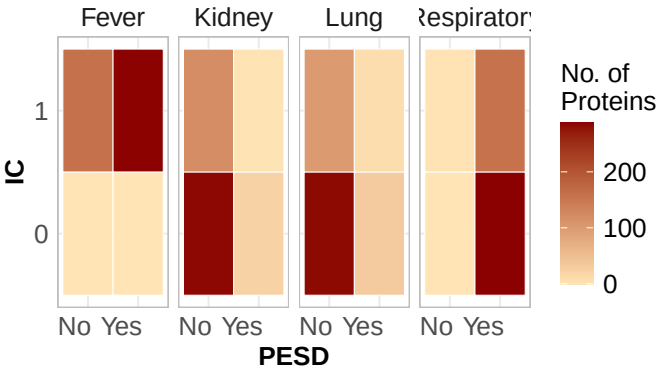

Figure 4. Significant proteins distribution.

4.3. Multi-Hot Encoding Results

The MHE transforms each of the 312 SP into a binary label and appends each evaluated parameter combination of a, p, and c as features, thereby completing the data preparation process for model training and evaluation. Tables 4a and 4b show the MHE transformation results for the SMOC1 and CD5 as an example, with a=1,p=3, c=1. For this cohort, the experiment found 1751 data points, and there were no empty labels for all combinations of a, p, and c.

Table 4. MHE transformation process for SMOC1 and CD5 proteins.

Protein	Adjusted p-value	FDR Threshold
SMOC1	1.26×10^{-44}	Significant
CD5	3.67×10^{-37}	Significant

(a) Respiratory Symptom=Yes, Immunocompromised

a	p	b	SMOC1	CD5
1	3	1	1	0
1	3	1	0	1

(b) MHE-transformed data

4.4. Training and Evaluation Results

The GSCV analysis combines different k-folds with *softmax* activation function, fixed dropout rate at 0.2 for h_1 and h_2, batch size, epochs, and n_1, n_2 parameters. Table 5 shows the best hyperparameter tuning during the GSCV process by testing different k-fold scenarios. The second scenario fixes the best found hyperparameters $n_1 = 32$, $n_2 = 12$, batch size = 5, and epochs = 50 and varies k-fold to obtain the F1-macro results, according to Table 6. Using more k-folds improves the F1-macro score.

Table 5. Cross-validation results for k-folds={20,30,40,50}

Parameter	Range	Result
n_1	{32,40,48}	**32**
n_2	{12,18,24}	**12**
Epochs	{50,100}	**50**
Batch Size	{5,10}	**5**
Activation	softmax	
Dropout rate	0.2	

Table 6. F1-macro variation per k-fold for the best hyperparameters found in Table 5

k-folds	Fits	F1-macro
20	720	0.72
30	1080	0.81
40	1440	0.86
50	1800	0.89

The evaluation metrics obtained in the experiment rely on the best hyperparameters found in Table 5, and follow the same pattern by variating the k-folds and testing each of the presented evaluation metrics. Table 7 shows the average metrics obtained in each k-fold scenario.

Table 7. Average metrics obtained in various k-fold scenarios

k-fold	Jaccard %	WD %	HL %	AUCROC %
10	44.04	0.63	0.35	71.81
20	71.97	0.64	0.32	85.96
30	81.31	0.64	0.32	90.70
40	85.99	0.64	0.32	92.94
50	**88.79**	**0.64**	**0.32**	**94.47**

For the best scenario in k-fold=50, the experiment identifies 4 labels with $J \geq 0.97$ as PB, according to Table 8. Each identified biomarker may contribute to cancer progression in the context of COVID-19. CEACAM3 regulates immune activation, particularly in neutrophils, and its dysregulation may promote immune evasion and tumor progression, exacerbated by COVID-19-induced inflammation [Skubitz 2024]. CNTN2 facilitates cancer cell migration and metastasis, potentially worsened by COVID-19-related neurological effects [Upadhyai et al. 2022]. NINJ1, involved in immune response and tissue repair, may accelerate metastasis and tissue damage under COVID-19-driven inflammation [Xu et al. 2022]. LPCAT2 supports cancer growth and metastasis, with COVID-

Table 8. Identified potential biomarkers

Tissue specificity	Protein	UniProt ID	Description
Bone marrow	NINJ1 CEACAM3	Q92982 P40198	Ninjurin-1 Carcinoembryonic cell adhesion molecule 3
Bone marrow, thyroid gland	LPCAT2	Q7L5N7	Lysophosphatidylcholine acyltransferase 2
Brain	CNTN2	Q02246	Contactin-2

19-related inflammation potentially intensifying its effects, particularly in lung cancers [Dahal et al. 2024].

5. Discussion

The DL multilabel classifier identifies four proteins from Table 8 with the highest Jaccard index ($J \geq 0.97$). Lowering this threshold increases the number of PB candidates but may introduce proteins less relevant to the possible relationship between cancer and COVID-19.

During the data preparation stage, each resultant data point contains only one label set to 1. Setting multiple labels to 1 for the same data point would lead the DL model to mathematically consider the combination of all labels set to 1 as the ground truth for training and evaluation. This scenario is equivalent to combining two or more non-correlated proteins as mutually dependent, potentially disrupting the preservation of proteomics characteristics for the cohort.

The application of the SKF technique effectively addresses the issue of imbalanced data within the cohort, as evidenced in the consistent performance metrics, including the Jaccard index, hamming loss, and Wasserstein distance, across the deep learning model labels. While the conventional confusion matrix may not be the most suitable approach for evaluating DL model performance, particularly in multilabel settings, our experiment bring the Jaccard index, AUCROC and Wasserstein distance as reliable tools for assessing the behavior of individual labels within the multilabel deep learning model. This, in turn, enhances the credibility of the identified PB. Furthermore, the complementary nature of the Wasserstein distance and hamming loss proves valuable in analyzing the losses during model training and evaluation.

6. Limitations

This section details the limitations encountered during the experiment and highlights the future directions of this work.

Firstly, with the available cohort of only 305 patients, the data preprocessing stage returns 1751 data points. In this scenario, we decide on the SKF technique to mitigate the imbalanced data during DL model training. For larger cohorts, we can instead adopt the traditional train-test split strategy for DL model training and evaluation, as more substantial cohorts would generate more data points. The MHE binarization strategy remains the same, preserving the overall methodology presented in this article. Secondly, the provided cohort brings blood draws for 1472 proteins only. The same architecture can reach

more than 5400 proteins, enhancing the scope of this methodology and the chances to find more PB. Lastly, we implement the entire MCM locally due to constraints in using a suitable existing BCP. Using a BCP makes the MCM scalable to parallelism and expands to Big Data scenarios for proteomic analysis.

7. Conclusion

This article presents a novel supervised deep learning multilabel classification framework to uncover PB associated with the interplay between cancer and COVID-19. The findings contribute to proteomics research by demonstrating promising evaluation metrics, and the identified candidate biomarkers offer valuable insights into this critical relationship. Implementing this methodology in a real-world biological clinical practice environment for larger-scale validation can enhance the results and benefit health organizations, laboratories, and academic institutions in the public health domain.

The MCM is a convenient tool for researchers, bioinformaticians, and hospital laboratory professionals seeking accelerated diagnosis and customized, guided therapy for oncology patients infected with COVID-19.

8. Acknowledgments

The authors would like to thank the Coordination for the Improvement of Higher Education Personnel - CAPES (Finance Code 001), and the National Council for Scientific and Technological Development - CNPq (Grant Numbers 307345/2023-8 and 404572/2021-9) for supporting this work.

References

Bai, Y., Yang, E., Han, B., Yang, Y., Li, J., Mao, Y., Niu, G., and Liu, T. (2021). Understanding and improving early stopping for learning with noisy labels. *Advances in Neural Information Processing Systems*, 34:24392–24403.

Boberg, E., Kadri, N., Hagey, D. W., Schwieler, L., El Andaloussi, S., Erhardt, S., Iacobaeus, E., and Le Blanc, K. (2023). Cognitive impairments correlate with increased central nervous system immune activation after allogeneic haematopoietic stem cell transplantation. *Leukemia*, 37(4):888–900.

Consortium, U. (2019). Uniprot: a worldwide hub of protein knowledge. *Nucleic acids research*, 47(D1):D506–D515.

Dahal, A., Hong, Y., Mathew, J. S., Geber, A., Eckl, S., Renner, S., Sailer, C. J., Ryan, A. T., Mir, S., Lim, K., et al. (2024). Platelet-activating factor (paf) promotes immunosuppressive neutrophil differentiation within tumors. *Proceedings of the National Academy of Sciences*, 121(35):e2406748121.

Doknic, A. and Möller, T. (2025). Mlmc: Interactive multi-label multi-classifier evaluation without confusion matrices. *arXiv preprint arXiv:2501.14460*.

Esti Anggraini, R. N., Machmudah, H., and Sarno, R. (2023). Hierarchical topic mining and multi-label classification on online news in bahasa. In *2023 International Conference on Advanced Mechatronics, Intelligent Manufacture and Industrial Automation (ICAMIMIA)*, pages 1–6.

Filbin, M. R., Mehta, A., Schneider, A. M., Kays, K. R., Guess, J. R., Gentili, M., Fenyves, B. G., Charland, N. C., Gonye, A. L., Gushterova, I., et al. (2021). Longitudinal proteomic analysis of severe covid-19 reveals survival-associated signatures, tissue-specific cell death, and cell-cell interactions. *Cell Reports Medicine*, 2(5):100287.

Fung, M. and Babik, J. M. (2021). Covid-19 in immunocompromised hosts: what we know so far. *Clinical Infectious Diseases*, 72(2):340–350.

García-Pedrajas, N. E., Cuevas-Muñoz, J. M., Cerruela-García, G., and de Haro-García, A. (2024). A thorough experimental comparison of multilabel methods for classification performance. *Pattern recognition*, page 110342.

Hossain, M. A., Rahman, M. Z., Bhuiyan, T., and Moni, M. A. (2024). Identification of biomarkers and molecular pathways implicated in smoking and covid-19 associated lung cancer using bioinformatics and machine learning approaches. *International Journal of Environmental Research and Public Health*, 21(11):1392.

Kluger, D. M. and Owen, A. B. (2024). A central limit theorem for the benjamini-hochberg false discovery proportion under a factor model. *Bernoulli*, 30(1):743–769.

Kocsmár, É., Kocsmár, I., Elamin, F., Pápai, L., Jakab, Á., Várkonyi, T., Glasz, T., Rácz, G., Pesti, A., Danics, K., et al. (2024). Autopsy findings in cancer patients infected with sars-cov-2 show a milder presentation of covid-19 compared to non-cancer patients. *GeroScience*, 46(6):6101–6114.

Li, B., Tang, X., Qi, X., Chen, Y., Li, C.-G., and Xiao, R. (2022). Emu: Effective multi-hot encoding net for lightweight scene text recognition with a large character set. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(8):5374–5385.

Liew, F., Efstathiou, C., Fontanella, S., Richardson, M., Saunders, R., Swieboda, D., Sidhu, J. K., Ascough, S., Moore, S. C., Mohamed, N., et al. (2024). Large-scale phenotyping of patients with long covid post-hospitalization reveals mechanistic subtypes of disease. *Nature immunology*, 25(4):607–621.

Liu, K., Qin, Z., Ge, Y., Bian, A., Xu, X., Wu, B., Xing, C., and Mao, H. (2023). Acute kidney injury in advanced lung cancer patients treated with pd-1 inhibitors: a single center observational study. *Journal of Cancer Research and Clinical Oncology*, 149(8):5061–5070.

Lv, C., Guo, W., Yin, X., Liu, L., Huang, X., Li, S., and Zhang, L. (2024). Innovative applications of artificial intelligence during the covid-19 pandemic. *Infectious Medicine*, page 100095.

Patel, M. A., Knauer, M. J., Nicholson, M., Daley, M., Van Nynatten, L. R., Cepinskas, G., and Fraser, D. D. (2023). Organ and cell-specific biomarkers of long-covid identified with targeted proteomics and machine learning. *Molecular Medicine*, 29(1):26.

Skubitz, K. M. (2024). The role of ceacam s in neutrophil function. *European Journal of Clinical Investigation*, 54:e14349.

Szeghalmy, S. and Fazekas, A. (2023). A comparative study of the use of stratified cross-validation and distribution-balanced stratified cross-validation in imbalanced learning. *Sensors*, 23(4):2333.

Upadhyai, P., Shenoy, P. U., Banjan, B., Albeshr, M. F., Mahboob, S., Manzoor, I., and Das, R. (2022). Exome-wide association study reveals host genetic variants likely associated with the severity of covid-19 in patients of european ancestry. *Life*, 12(9):1300.

Wik, L., Nordberg, N., Broberg, J., Björkesten, J., Assarsson, E., Henriksson, S., Grundberg, I., Pettersson, E., Westerberg, C., Liljeroth, E., et al. (2021). Proximity extension assay in combination with next-generation sequencing for high-throughput proteome-wide analysis. *Molecular & Cellular Proteomics*, 20.

Xu, Q., Yang, Y., Zhang, X., and Cai, J. J. (2022). Association of pyroptosis and severeness of covid-19 as revealed by integrated single-cell transcriptome data analysis. *ImmunoInformatics*, 6:100013.

Yadalam, P. K., Arumuganainar, D., Natarajan, P. M., and Ardila, C. M. (2025). Predicting the hub interactome of covid-19 and oral squamous cell carcinoma: uncovering aldh-mediated wnt/β-catenin pathway activation via salivary inflammatory proteins. *Scientific Reports*, 15(1):4068.

Zhou, J., Lakhani, I., Chou, O., Leung, K. S. K., Lee, T. T. L., Wong, M. V., Li, Z., Wai, A. K. C., Chang, C., Wong, I. C. K., et al. (2023). Clinical characteristics, risk factors and outcomes of cancer patients with covid-19: A population-based study. *Cancer Medicine*, 12(1):287–296.