PLOS ONE

RESEARCH ARTICLE

Predicting mechanical ventilation effects on six human tissue transcriptomes

Judith Somekh*, Nir Lotan, Ehud Sussman, Gur Arye Yehuda

Department of Information Systems, University of Haifa, Haifa, Israel

* judith_somekh@is.haifa.ac.il

OPEN ACCESS

Citation: Somekh J, Lotan N, Sussman E, Yehuda GA (2022) Predicting mechanical ventilation effects on six human tissue transcriptomes. PLoS ONE 17(3): e0264919. https://doi.org/10.1371/journal.pone.0264919

Editor: Serdar Bozdag, University of North Texas, UNITED STATES

Received: February 13, 2021

Accepted: February 21, 2022

Published: March 10, 2022

Data Availability Statement: The GTEx data is available for download from (https://www.gtexportal.org/home/datasets).

Funding: The research was funded by a grant from the Data Science Research Center (DSRC) at the University of Haifa and by an internal funding from the University of Haifa. There was no additional external funding received for this study.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Background

Mechanical ventilation (MV) is a lifesaving therapy used for patients with respiratory failure. Nevertheless, MV is associated with numerous complications and increased mortality. The aim of this study is to define the effects of MV on gene expression of direct and peripheral human tissues.

Methods

Classification models were applied to Genotype-Tissue Expression Project (GTEx) gene expression data of six representative tissues–liver, adipose, skin, nerve-tibial, muscle and lung, for performance comparison and feature analysis. We utilized 18 prediction models using the Random Forest (RF), XGBoost (eXtreme Gradient Boosting) decision tree and ANN (Artificial Neural Network) methods to classify ventilation and non-ventilation samples and to compare their prediction performance for the six tissues. In the model comparison, the AUC (area under receiver operating curve), accuracy, precision, recall, and F1 score were used to evaluate the predictive performance of each model. We then conducted feature analysis per each tissue to detect MV marker genes followed by pathway enrichment analysis for these genes.

Results

XGBoost outperformed the other methods and predicted samples had undergone MV with an average accuracy for the six tissues of 0.951 and average AUC of 0.945. The feature analysis detected a combination of MV marker genes per each tested tissue, some common across several tissues. MV marker genes were mainly related to inflammation and fibrosis as well as cell development and movement regulation. The MV marker genes were significantly enriched in inflammatory and viral pathways.

Conclusion

The XGBoost method demonstrated clear enhanced performance and feature analysis compared to the other models. XGBoost was helpful in detecting the tissue-specific marker genes for identifying transcriptomic changes related to MV. Our results show that MV is associated with reduced development and movement in the tissues and higher inflammation

and injury not only in direct tissues such as the lungs but also in peripheral tissues and thus should be carefully considered before being implemented.

Introduction

Mechanical ventilation (MV) is a lifesaving intervention used for patients with respiratory failure. Although its therapeutic effects are well known, MV is also associated with numerous complications including significantly higher infection rates and lung injury, which prolong the duration of time spent on the ventilator and increase mortality [1], which can exceed 24% of those on ventilators [2]. The 2019 COVID-19 outbreak, resulting in MV treatment for numerous patients, has moved the question regarding invasive ventilation usage [3] to center stage.

Clinical and transcriptomic effects of MV on tissues

The connection between treatment with invasive MV and pulmonary infections is so pervasive that the terms Ventilator-Associated Pneumonia (VAP) and ventilator-associated events (VAE) [2] were coined. VAE includes all the complications related to mechanical ventilation, broadening the horizon of possible consequences beyond those infection-related [4–6]. An example of VAE is the weaking of the diaphragm muscles [7, 8] that were observed after only twelve hours on an mechanical ventilator [5]. A large-scale study of 549 patients showed that samples of patients on MV exhibited increased lung inflammation and injury by testing inflammatory markers related to lung injury derived from blood samples [9]. MV contributes to mortality by inducing an inflammatory response in the lungs similar to that observed in acute respiratory distress syndrome (ARDS) [4], which can lead to multisystem organ failure. Although the most obvious clinical abnormalities in ALI (acute lung injury)/ARDS are related to the lungs, the most common cause of death is not due to hypoxia but to multiple organ dysfunction syndrome (MODS) [10]. Indeed, MV has been associated with greater risk of kidney failure [11], and diminished neurocognitive function in the brain [12]. Pinhu et al. [13] suggested two possible mechanisms through which MV induces multiple organ failure that are related to VAP and lung injuries that reduce the rate of organ perfusion.

Better understanding of the pathophysiology leading to the development of MODS in patients on MV should help in the development of approaches to interrupt the cascades leading to the syndrome [10]. The exact molecular mechanics of how MV affects peripheral organs is less obvious and the understanding of gene expression alteration in patients on MV may help. The ramifications of MV on human tissues' gene expression are poorly explored and have focused mainly on the lungs [2, 14, 15]. MV was shown to stimulate the expression of the SARS-Cov-2 receptor ACE2 in the lungs [14] and a large-scale genomic research explored the effects of MV on the lung transcriptome [2, 14, 15]. The work in [15] used the GTEx gene expression data to detect several distinct gene expression clusters in the lungs, including a large cluster of genes associated with type II pneumocytes related to cells that proliferate in ventilator associated lung injury. Human and animal studies have demonstrated that MV using large tidal volumes ($\geq$12 ml/kg) induces a potent inflammatory response and can cause acute lung injury [2]. Using a mouse model, [2] showed that non-injurious MV on its own initiates a proinflammatory transcriptional program in the lung. They compared breathing mice and mice on non-injurious MV (tidal volume of 10 ml/kg) and undertook an unbiased approach to partially decipher the complex network of related pathways. They showed that the low tidal volume still activates a transcription program of severe lung injury. In their previous

work [16], they showed that the combination of non-injurious MV and low-dose exposure to bacterial products can cause severe lung injury in mice, implying a comodulatory role for MV in lungs that are at risk. In another research attempting to decrease lung injury, MV was shown to affect genes expression of the lung areas [17]. In summary, while MV directly affects the lungs, investigating specific mechanistic effects on peripheral non-direct tissues has only been partially explored [10–12]. In addition, a large-scale investigation of MV-related changes in gene expression and molecular pathways in peripheral tissues is lacking.

Machine learning methods for predicting MV and mortality

Machine learning models have been widely used in medical applications, including to predict MV and mortality. For example, several machine learning methods were used for predicting mortality of patients with acute kidney injury hospitalized in the ICU (Intensive Care Unit) [18]. Artificial Neural Networks (ANNs or NNs) were used on data of breathing patterns to predict asynchronous breathing (AB) during MV [19]. Machine learning models used clinical data to predict MV mortality of COVID-19 patients in emergency rooms and in-hospital once the patient was admitted [3].

Machine learning approaches have been used not only to predict but also to detect the set of features that drive the prediction, e.g., analysis of gene expression data to discover novel marker genes, gene signatures and related pathways and networks, to differentiate between conditions. Machine learning can yield a list of differential genes that combine to drive the prediction and consider the dependence between the genes. For example, Grunwell et al. [20] applied machine learning to nanostring transcriptomics on primary airway cells and a neutrophil reporter assay to discover gene networks differentiating pediatric acute respiratory distress syndrome from non-pediatric ARDS. Cai et al. [21] used gene expression data fed into three modeling methods—logistic regression, random forest and neural network—to develop a diagnostic gene signature for the diagnosis of VAP.

The Genotype-Tissue Expression (GTEx) [22] Project is a comprehensive public resource that includes tissue-specific gene expression data from nearly 1000 relatively healthy post-mortem and some material from "normal" surgical specimen human donors. The GTEx data of several tissues were successfully used in developing a machine learning model to predict the time since death of the donors [23].

In this study, we analyzed the GTEx [22] RNA-sequencing data to investigate the impact of MV on the transcriptomes of six representative human tissues. The study objective was to define gene signatures and pathways differentially expressed in direct and peripheral tissues of patients on MV. We constructed three machine learning models for each tissue and analyzed the important features to predict patients who were on MV and those who were not and to detect the significant MV-related gene signatures, using the analysis of the transcriptomes. Specifically, our models predict whether the donor was subject to ventilation prior to death. We trained and evaluated models on gene expression data of hundreds of samples for each tissue and thousands of genes as features.

To the best of our knowledge, ours is the first study to describe an investigation of the genomic effects of MV on multiple peripheral tissues using machine learning algorithms and enrichment analysis of genes across tissues in human donors using gene expression data.

Methods

Data preprocessing

GTEx RNA-Seq data of 54 human tissues and 17382 RNA-seq samples from nearly 1000 donors was downloaded from the GTEx database (https://www.gtexportal.org/home/datasets,

v8), and their transcript per million (TPM) values were log2-transformed. 18,680 protein-coding genes were retained. Outlier samples were filtered and all genes within each tissue were quantile normalized (to remove background and sample effects). For outlier removal, for each tissue we removed outlier samples by applying an Isolation Forest algorithm [24], with a parameter of 0.01. Accordingly, the 1% of the remaining samples that had the highest levels of variation from the other samples were also removed. Genes with zero variance were excluded from the calculation (e.g., for adipose–subcutaneous, 262 genes were excluded). For a given tissue, genes having at least 0.1 TPM in 80% or more of the samples were retained. S1 Table in S1 File presents the number of samples and features per tested tissue type after preprocessing and correcting for confounding factors. For example, for adipose–subcutaneous, there are 544 samples; for each sample, there are 16,052 genes that we use as features in the machine learning models.

Confounding factor adjustment

Our previous work [25] showed that linear regression-based adjustment of the heterogenous GTEx data outperforms other methods in preserving the biological signal—which is relevant here. Thus, we used linear regression models to adjust for the known confounding factors—experimental batch, ischemic time (time elapsed between actual death and sample extraction), gender and age. Age covers the 20–80-year range and is partitioned into 10-year intervals (embedded in the GTEx dataset).

The type of death classification of the samples (DTHHRDY = death circumstances) is based on a four-point Hardy Scale. 0 represents cases on mechanical ventilator prior to death, 1 and 2 represent non-ventilation deaths (short and intermediate duration prior to death) of healthy individuals, and 3 and 4 represent non-healthy individuals. As the focus of the research is on healthy individuals at the time of death, we excluded samples with a DTHHRDY value of 3 "Intermediate death for ill patients" and 4 "Slow death" that represent non-healthy individuals with a long-term illness (these also comprised a small number of samples).

We aggregated the DTHHRDY into two categories—death type 0 (ventilation), a subject who was on a ventilation machine prior to death and death type 1 (non-ventilation), a subject who was not connected to a ventilator at the time of death. Ischemic time is the time in minutes that elapsed between death and sample extraction. We found that there was a correlation between the ischemic time (SMTSISCH) and DTHHRDY (ventilation vs. non-ventilation) that wrongly skewed the ischemic time coefficient when we used both as predictors in the linear regression model (see explanations and plots in S4 and S5 Figs in S1 File). This phenomenon is the result of the fact that individuals on MV were already in the hospital and this resulted in shorter time (ischemic time) between time of death and sample collection. This phenomenon was detected previously [23] and dealt with by correcting for only one of these two confounding factors. Here we developed an improved approach to correct the data for ischemic time but with minimal harm to the ventilation signal. We performed linear regression in two steps. We first corrected for age, sex and batch and then for ischemic time, by inferring its coefficient for each group separately; thus, we did not skew by the correlated death type. We used the averaged coefficients calculated for each group (ventilation/non-ventilation) independently as explained below. After correcting the data with our linear regression model, we used the residuals as the expression values for further analysis.

The two-step process of extracting residuals was as follows:

$$Residual0_i^j = Exp_i^j - \sum_{n=1}^{N} Coef_{i,n} \times Confounder_n^j \quad (1)$$

Exp_i^j is the expression level of gene *i* in sample *j*. $Coef_{i,n}$ is the multiple linear regression coefficient gene *i* in coefficient *n*. $Confounder_n^j$ is the phenotype confounding value for sample j and confounding factor n. $ResidualL0_i^j$ is the residual level of the value for gene i and sample j. The confounding factors in the model and their corresponding confounding coefficients were gender, age, and the experimental batch. Then, the ischemic rate coefficient was calculated for each two ventilation types separately in order to correct independently for ischemic time and not for ventilation type (which are correlated). We separated the residuals by the two ventilation types (ventilation and non-ventilation). An unweighted average of the coefficients by ventilation type $Ischem_i$ per gene, was taken. The residual used was:

$$Residual_i^j = ResidualL0_i^j - min(0, Ischem_i) \times Ischemictime_j \quad (2)$$

$Ischem_i$ is the average of the factor calculated for ischemic time for death type 0 (ventilator) and death type 1 (non-ventilator). If the average of the ischemic coefficients was greater than 0, then there was no further adjustment performed for ischemic time. $Ischemictime_j$ is the ischemic time for sample j. $Residual_i^j$ is the residual value once the two-step regression process was completed for gene i and sample j. Ideally, we would have liked to perform the regression in one step; however, we found that for the genes with the highest predictive rates for ventilation type, this would have yielded skewed ischemic time coefficients. Further details and explanatory plots of the correction approaches we tested and the final two-step linear regression approach we used can be found in S1 File.

Machine learning methods

We selected three different algorithms and six main tissues and built machine learning classifiers for each tissue—resulting in a total of 18 machine learning models. Each model is a binary classifier designed to predict the ventilation/non-ventilation samples based on the RNA-seq gene levels per samples, i.e., the features, derived from the specific tissue. The selection of algorithms was somewhat limited given the nature of the data: since ours was a relatively small sample size (~200–700 items) having a large number of features (~13-16k), not all machine learning methods were expected to achieve good results. Due to the relatively small sample size, we decided to experiment with two different decision tree-based algorithms (described bellow): Random Forest (RF), and a boosted decision tree, the eXtreme Gradient Boosting technique (XGBoost) [26]. We expected that the boosted decision trees would be very effective in our context. Tree boosting machines are a family of powerful machine learning techniques that have shown considerable success in a wide range of practical applications. An additional advantage of tree boosting machines is the explainability capabilities of these models, which can help in validating the correctness of the model, by checking the relevance of the most significant gene levels to the tested conditions, i.e., ventilation vs. non-ventilation, and learning the biological signs that the model is detecting. In addition, and given the popularity of using Deep Learning and the extensive usage of it in many different use cases these days, we also chose to use the Artificial Neural Network (ANN) [27] as one of our methods, although it may be less effective when working with tabular data and a relatively small sample size.

Random forest

RF is an ensemble algorithm that combines multiple decorrelated decision tree prediction variables based on each subset of data samples [28]. RF has been extremely successful as a general-purpose classification and regression method, proven to be a computationally efficient

technique that can operate quickly over large datasets, and is easy to implement. RF can also handle a large number of input variables without overfitting [28]. In general, the RF approach creates several randomized decision trees, then combines and aggregates their predictions by averaging.

The RF model was built using the scikit-learn [29] Random Forest Regressor. We used a model configuration similar to the XGBoost (presented below), i.e., we use 100 estimators with a maximum depth of 3 and a learning rate of 0.1.

Tree boosting—XGBoost

Boosting [30] is a commonly used machine learning method that attempts to improve the accuracy of a given learning algorithm. Boosting is done by creating an ensemble learner from several weaker models. The ensemble takes a set of predictors, all aiming to predict the same target, and combines them together to form a stronger predictor. Friedman [31] was the first to propose a gradient-descent-based formulation of boosting, a method that improves the approximation accuracy.

The XGBoost technique [26] method is based on Friedman's gradient boosting but introduces additional improvements that increase the technique's performance and the accuracy of its results. While in the original gradient boosting model, the trees are built in series, XGBoost does this in a parallel way, similar to the RF method that grows trees in parallel to each other, and each tree tries to compensate for the areas in which the previous tree was less accurate. This method also uses regularization terms to control the variance of the fit and control the flexibility of the learning task, while obtaining models that generalize better to unseen data. XGBoost has been extensively used recently and shown useful in solving different real-world problems, for example, pathway analysis of biomedical data [32] and diagnosis of chronic kidney disease [33].

We executed the XGBoost algorithm to create a set of 100 decision trees for each tissue. An example of one generated XGBoost tree and its branches for muscle-skeletal tissue classification is provided in S1 Fig (S1 File). Similarly, S2 Fig in S1 File. provides one of the tree branches that were generated for the adipose-subcutaneous tissue. After the tree is created, it can be used for prediction when each tree node, depicted by an ellipse as seen in S1 and S2 Figs (S1 File), that represents a condition on a gene expression value is checked for each predicted sample. For each tree node from the top of the tree, if the sample's value equals the tree node specification, the selected path in the decision tree is the 'yes' path. Else the 'no' path is selected. If the value is missing, the 'missing' path is selected. For example, for the provided "Adipose–Subcutaneous" tree in S1 Fig (S1 File), the level of MXRA5 in the sample is compared to 0.32043466. If the value is smaller than 0.32043466, or there is no measurement, the left path is selected. Eventually the selected sample receives a score (for being ventilation/non-ventilation) for each tree. The scores of 100 trees will be combined to determine the selected ventilation/non-ventilation class. Each of the 100 trees may contain different genes and check different values for these gene levels.

To develop the XGBoost binary classifier we used the XGBoost Python library [34]. To avoid overfitting, while maintaining high performance predictions, we configured the XGBoost to use 100 estimators with a maximum depth of 3. We found that a high number of estimators (100) performed better than lower values (e.g., 30 or 50). A learning rate of 0.1 was found to be effective in this case (lower values were tested). For the rest of the parameters, the default values were used.

Artificial Neural Network (ANN)

ANN [27, 35] is a computational model inspired by biological models, which often exceed the performance of previous forms of artificial intelligence used in many common machine learning tasks. It is basically a data processing system composed of a combination of simple, interconnected processing elements, in a predesigned architecture. The ANN basic building block is a simple mathematical function that includes three steps: multiplication, summation and activation. For the purpose of this research, we created a network with a shallow topology of one hidden layer, as shown in S6 Fig (S1 File). Additional topologies were tested, but we found that a single hidden layer network provides the best results. The selected topology includes an input layer with the dimensions of the number of features (based on the relevant organ), a hidden layer of 100 nodes with a Rectified Linear Unit (ReLU) activation function [36], and a single output layer with a sigmoid activation function. We use binary cross-entropy as the loss function, and an adaptive moment estimation optimizer (Adam) [35]. The neural network model was designed using the Keras [37] network, with TensorFlow [38] as its backend.

To confirm the effectiveness of the XGBoost model in predicting MV, we used ANN and RF, widely used machine learning models, for comparison and summarized the advantages and disadvantages of each model in Table 1.

Computation resources and frameworks

The training of the models was conducted on an Intel(R) Core(TM) i9-7920X CPU @ 2.90GHz computer with 24 CPUs, with 128GB RAM and an NVIDIA Corporation GV100 [TITAN V] (rev a1) GPU. Training in this setup lasted about five days.

10-fold cross-validation

For model performance evaluations, i.e., to evaluate whether the model is accurate and not overfitted and due to the relatively low number of samples (~200–700 samples), we used K-Fold Cross Validation—more precisely, the scikit-learn [39] implementation of Stratified K-Fold Cross Validation without shuffling. This cross-validation is a variation of K-Fold that

Table 1. Comparison between the machine learning models used in the study.

Model	Advantages	Disadvantages
XGBoost	• Effective for a relatively small number of samples with a large number of features	• May exhibit overfitting if hyperparameters are not adjusted correctly
	• Encapsulated explainability capabilities that can help validate the correctness of the model, e.g., by checking the relevance of the most significant gene levels to the tested condition	• Applicable for numeric features only
	• Includes improvements to the original gradient boosting model that increase the performance and the accuracy of the results	
RF	• Easy to implement both for classification and regression tasks	• Lower performance than more modern methods
	• Provides some level of explainability	• Nonoptimal performance when classes are unbalanced
	• Avoids overfitting	
ANN	• Excels at cognitive tasks (image/video/text/voice data)	• Requires a large number of samples
		• Hard to explain and detect the feature importance
		• May be less effective with tabular data

https://doi.org/10.1371/journal.pone.0264919.t001

returns stratified folds. The folds are made by preserving the percentage of samples for each class. We chose a split of 10, so each time 90% of the data is used for training, and 10% for validation. We used the same fold and data when building the ANN, RF and XGBoost models to assure we compared the models under exactly the same terms.

For each organ we configured each model to use the organ's gene levels as features, with the ventilation/non-ventilation types as the predicted target. We performed 18 experiments, each per distinct model and tissue. Each experiment included a 10-fold cross-validation prediction, and we saved the results of the 10 executions. We averaged the results of the 10 executions as well as the feature importance. For each prediction we tracked the average and standard deviation of the Area Under the ROC curve (AUC), accuracy, F1 score, recall and precision of each experiment.

Feature analysis

Feature analysis was performed using Lundberg's approach for explaining boosted trees called SHAP—SHapley Additive exPlanations [40]. SHAP offers a high-speed precise algorithm that can explain the output of any machine learning model, in particular tree ensemble methods. SHAP calculates values for each feature, representing how much each feature contributes to push the model output from the base value (the average model output over the training dataset we processed) to the model output.

Pathway enrichment analysis

We imported the full list of the marker genes used for the predictions into the web-based Enricher tool [41] using KEGG (Kyoto Encyclopedia of Genes and Genomes) 2021 pathways.

Results

We used RNA-seq gene expression data from the GTEx project [22] for six representative human tissues—adipose-subcutaneous, liver, lung, muscle-skeletal, nerve-tibial and skin-sun exposed (lower leg) (see S1 Table in S1 File for the full number of samples and genes for each tissue). We aggregated the samples into ventilation and non-ventilation groups as described in the Methods section. The tissues' gene expression data was preprocessed and corrected for confounding factors as described in the Methods section and in the S1 File.

To detect a combination of marker genes signifying MV usage, we created 18 prediction models using the RF, XGBoost decision tree and ANN methods to classify the ventilation and non-ventilation samples for the six tissues. To evaluate the predictive performance of each model, we compared the classifiers using the AUC, accuracy, precision, recall and F1 scores. We finally conducted feature analysis per each tissue to detect MV marker genes, followed by pathway enrichment analysis for these genes.

Classification comparison

Tables 1 and 2 provide a detailed comparison between the methods performance in means of the different metrics: AUC, Accuracy, F1 Score, Recall and Precision. The numbers provided here are the average of the 10-fold executions across the six tissues. Detailed results including standard deviation between executions can be found in S2 Table (S1 File). It is easy to see here and in Fig 1 that XGBoost outperforms the ANN and RF models in all metrics aside from recall, in which the ANN model outperforms XGBoost but only by a small margin. We added the AUC metric in our analyses (see Table 2) since the class distribution within the data is proportional but not fully balanced (see S1 Table in S1 File for the number of samples in each class).

Table 2. Binary classification model evaluations.

Model \ Evaluation	Average Accuracy	Average F1_score	Average Recall	Average Precision
Neural Net	0.934	0.916	**0.931**	0.903
Random Forest	0.912	0.880	0.862	0.905
XGBoost	**0.951**	**0.936**	0.924	**0.951**

https://doi.org/10.1371/journal.pone.0264919.t002

Table 3 provides a comparison of the results of the 18 models' performance, focusing on the AUC and providing the average AUC and AUC per organ. Clearly, even when looking at each tissue individually, XGBoost outperforms ANN and RF, although in some cases the difference is not significant. Given that XGBoost outperforms the two other methods (see also Fig 1), while providing good feature analysis capabilities, from this point on we focus on the modeling, experiment and results of the XGBoost model.

Tissue-specific MV marker genes

To detect the most predictive and ventilation discriminant genes, for each tissue we performed feature analysis using SHAP [40] to explain the boosted trees. SHAP calculates values for each

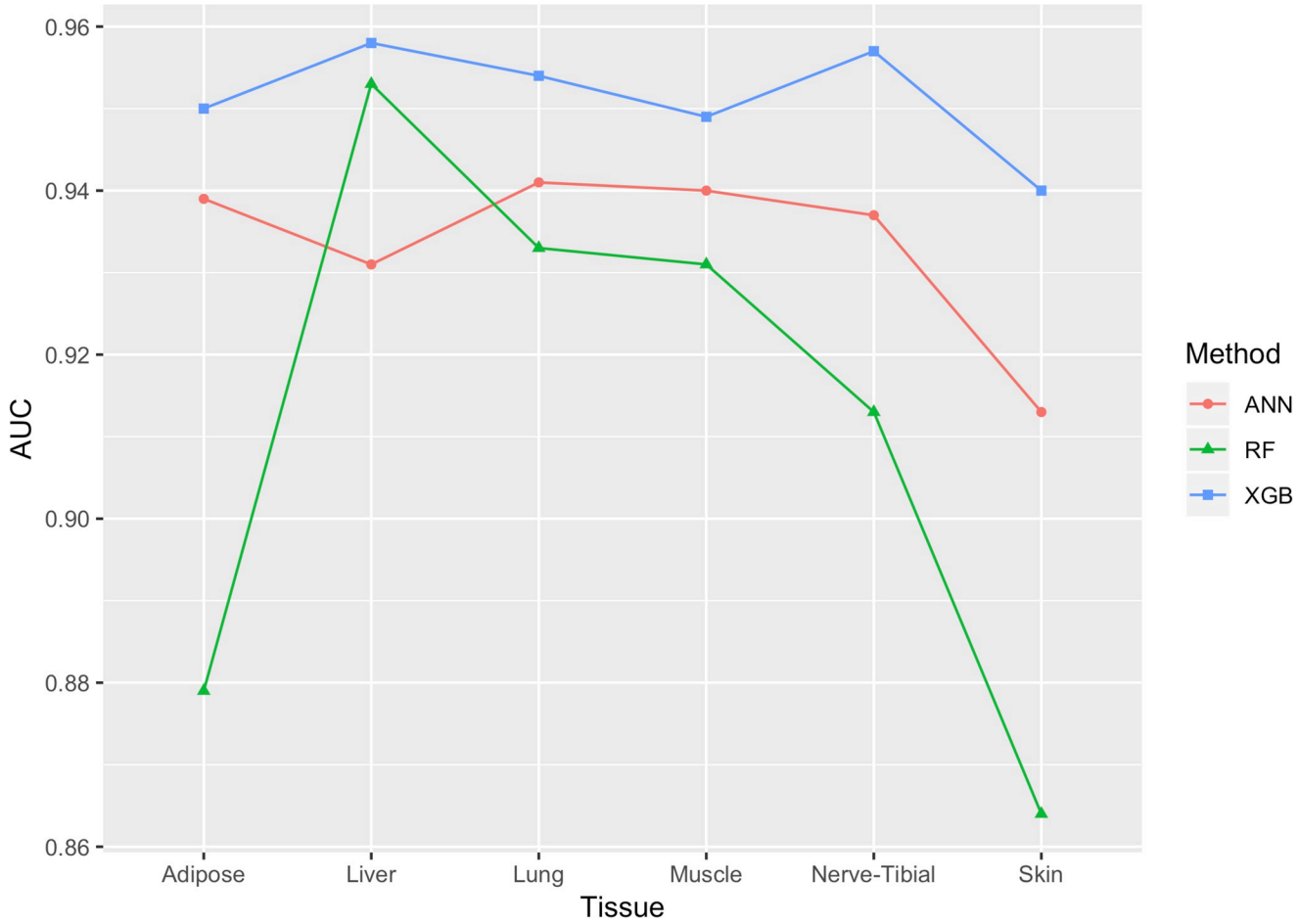

Fig 1. AUC comparison of the 18 classifiers. It can be seen that the XGBoost model outperforms the RF and ANN models for the six tested tissues.

https://doi.org/10.1371/journal.pone.0264919.g001

Table 3. Binary classification models' AUCs for the different organs.

Model \ AUC per organ	Average AUC	Adipose-Subcutaneous	Liver	Lung	Muscle-Skeletal	Nerve-Tibial	Skin-Sun Exposed (Lower leg)
Neural Net	0.934	0.939	0.931	0.941	0.940	0.937	0.913
Random Forest	0.912	0.879	0.953	0.933	0.931	0.913	0.864
XGBoost	**0.951**	**0.950**	**0.958**	**0.954**	**0.949**	**0.957**	**0.940**

https://doi.org/10.1371/journal.pone.0264919.t003

feature, representing how much each feature contributes to elevate the base value of the model, the average model output over the training dataset we processed. Fig 2A and 2B demonstrates the feature analysis results for lung and muscle-skeletal respectively and presents the mean absolute value of the SHAP values for each feature in our model. The top 20 features' average SHAP impact on model output magnitude in absolute values are presented. A summary of SHAP scores for the top 20 genes in all tested tissues are provided in Table 4.

The SHAP approach can also explain how each feature (gene) contributes to the classification per class. Fig 3 provides a plot of genes sorted in descending order by gene importance, representing the sum of SHAP value magnitudes over all samples and uses SHAP values to show the distribution of feature impacts on the model output. The horizontal location shows the impact of each feature, i.e., whether the effect of that value is associated with a higher or lower prediction. The colour relates to the original values of each gene across samples and shows whether that variable is high (in red) or low (in blue) for that observation. Red represents a higher value of the gene for the ventilation samples compared to the average values across all samples in the measured tissue; blue represents a low measured value. The x-axis in Fig 3 is the SHAP values of each gene and represents the impact of the gene on model output

Gene Name	Feature Importance
PHF13	0.296
MT4	0.295
HRG	0.237
TBC1D22B	0.220
LCE5A	0.185
ALPK3	0.184
MXRA5	0.168
LCE2C	0.158
CYP1A2	0.146
ADRA1B	0.142
TERF2IP	0.128
LCE2A	0.114
SPINT3	0.110
B3GNT2	0.107
SST	0.099
DPRX	0.095
DYRK2	0.093
AGTR2	0.092
PIP	0.092
LSM11	0.089

A

Gene Name	Feature Importance
GKN1	0.523
CLEC3B	0.339
TET1	0.270
FGF6	0.229
SMCO1	0.184
CCND1	0.173
ZCCHC24	0.164
SPSB4	0.154
EGR1	0.153
TFF1	0.145
KRT6B	0.142
CEACAM6	0.125
TLL2	0.123
FITM1	0.120
MRPL16	0.117
SPRR2A	0.108
SPATA25	0.101
TBC1D12	0.099
SLN	0.096
SOCS4	0.090

B

Fig 2. Top 20 genes and average SHAP impact (absolute values) on the magnitude of model classification output. (A) Values for the lung tissue. (B) Values for muscle-skeletal tissue. It can be seen that GKN1 gene expression values have the highest impact on the MV classification.

https://doi.org/10.1371/journal.pone.0264919.g002

Table 4. Top 20 genes with highest importance SHAP scores in each tissue.

	Adi-pose	Score	Liver	Score	Lung	Score	Muscle	Score	Nerve Tibial	Score	Skin	Score
1	**MXRA5**	0.49	MGMT	0.63	PHF13	0.30	**GKN1**	0.52	FDCSP	0.54	AVP	0.53
2	VENTX	0.42	DNAJB4	0.55	MT4	0.29	**CLEC3B**	0.34	HSD11B2	0.50	**TFF1**	0.41
3	**EGR1**	0.31	C5orf24	0.26	HRG	0.23	TET1	0.26	CRABP1	0.44	ODF3L1	0.24
4	PELO	0.27	TOR1A	0.24	TBC1D22B	0.23	FGF6	0.21	**EGR1**	0.31	**MXRA5**	0.23
5	**CLEC3B**	0.25	GATAD1	0.22	LCE5A	0.20	SMCO1	0.19	SLN	0.22	**KRT**1	0.23
6	**KRT**6C	0.22	C12orf60	0.14	ALPK3	0.18	SPSB4	0.18	SLITRK6	0.18	**CYS1**	0.20
7	**CYS1**	0.20	GPRIN1	0.13	**MXRA5**	0.17	ZCCHC24	0.17	**KRT**20	0.18	NEURL2	0.19
8	MUC21	0.20	KCNJ8	0.13	LCE2C	0.16	**KRT**6B	0.17	XIRP1	0.15	**RP11**-676J12.7	0.19
9	C22orf31	0.19	DRD4	0.12	CYP1A2	0.14	CCND1	0.16	HBQ1	0.14	**KRT**AP5-6	0.17
10	CX3CL1	0.17	ENTPD7	0.12	ADRA1B	0.14	**EGR1**	0.15	BHLHE40	0.13	NIPSNAP3A	0.15
11	C10orf99	0.15	SLC25A21-AS1	0.11	TERF2IP	0.12	**TFF1**	0.14	SFRP2	0.13	FLG	0.13
12	DEFA6	0.15	OPN1SW	0.10	LCE2A	0.12	FITM1	0.13	GUCA2A	0.12	RD3	0.11
13	PRND	0.14	PAQR8	0.09	SST	0.11	MRPL16	0.13	CD248	0.12	**EGR3**	0.10
14	SMCP	0.13	STX11	0.09	B3GNT2	0.10	TLL2	0.12	**RP11**-10J21.3	0.12	ATP4B	0.10
15	TNFRSF21	0.09	LRRC40	0.08	DYRK2	0.10	SPRR2A	0.12	DKK4.00	0.09	WFDC12	0.10
16	**GKN1**	0.09	CECR6	0.08	DPRX	0.10	CEACAM6	0.12	**RP11**-268J15.5	0.09	OCM2	0.10
17	HUS1B	0.09	OMG	0.08	CHP2	0.09	SLN	0.11	DUPD1	0.09	DUSP1	0.10
18	TIFAB	0.09	IFITM10	0.08	AGTR2	0.09	SPATA25	0.10	TAS2R46	0.08	SCNN1G	0.09
19	P2RY13	0.08	TMEM200C	0.07	SPINT3	0.09	SOCS4	0.09	DBX2	0.08	LGALS7B	0.09
20	TCF21	0.08	PSPN	0.07	LSM11	0.09	TBC1D12	0.09	PSD2	0.08	IER2	0.09

Genes that are common to more than two tissues are highlighted in bold.

https://doi.org/10.1371/journal.pone.0264919.t004

in muscle-skeletal (Fig 3A) and lung (Fig 3B) tissues. A low SHAP value means that this sample is more likely to be a ventilation sample and a high SHAP value means that this sample is more likely to be a non-ventilation sample. This approach reveals, for example, that for most cases, a high value of CLEC3B (the second gene from the top in Fig 3A) has a high and positive impact on non-ventilation prediction and increases the chances of the sample to be classified as taken from a non-ventilator group. The high comes from the red color and the positive impact shown on the x-axis. The CLEC3B gene has been reported to regulate muscle development [42]. An example of negative correlation is the AGTR2 gene (fourth from the bottom in Fig 3B) that has a high and negative impact on the non-ventilation prediction, i.e., it is lower in the non-ventilation group and higher in the ventilation group. We note that the angiotensin II receptor 2 (AGTR2) gene [43] is expressed in lung fibrosis. The diagram also illustrates the importance of using multiple features and feature combinations for achieving the high accuracy. We note that the tree-based method is based on calculating threshold values for a combination of genes to drive the prediction. An example of such combination, an XGBoost tree branch, is presented in S1 and S2 Figs (S1 File) and explained in the Methods section. One feature is not enough to differentiate between the samples and a combination of features is required. It is easy to see, for example, that for muscle-skeletal, high values of GKN1 correlate in most cases to the non-ventilator group. This gene's levels, however, are not good enough to be used as a single discriminator, since there are several cases of samples in the ventilator group that have high levels of GKN1. Only by using the combination of gene levels can the model achieve high accuracy/AUC. Additional histograms illustrating the differences in CLEC3B and AGTR2 gene values in muscle-skeletal and lung tissues for the ventilation and non-ventilation types are included in S3 Fig (S1 File).

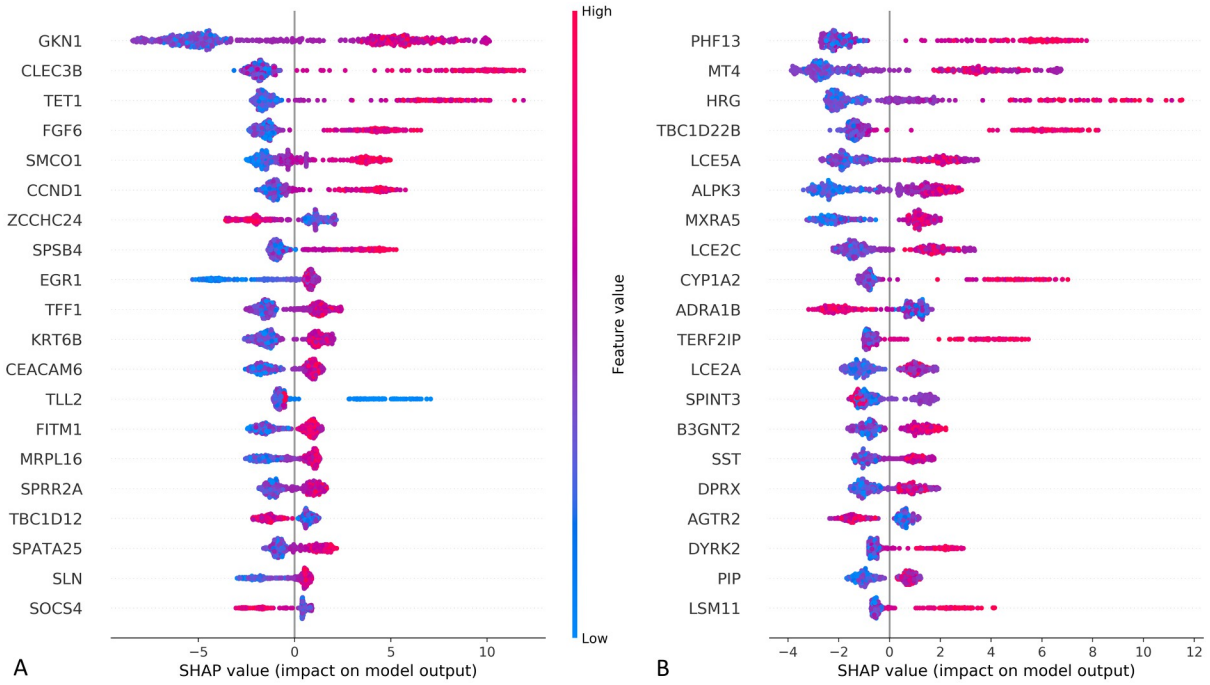

Fig 3. SHAP variable importance plots. (A) SHAP values for muscle-skeletal tissue. (B) SHAP values for the lung tissue. The plot includes all samples in the training data and the values represent the impact of the gene on model prediction output. SHAP values explain to what extent the feature (gene) contributes to the prediction of the model.

https://doi.org/10.1371/journal.pone.0264919.g003

Table 4 summarizes the top 20 most predictive genes for the six tissues. Genes common to several tissues appear in bold.

We used the web-based Enricher tool [41] for pathway enrichment analysis of all the genes per tissue that participated in the predictions, using KEGG 2021 Human pathways. We tested the enrichment of 1005 muscle-skeletal genes, 971 adipose-subcutaneous genes, 883 nerve tibial genes, 1089 skin genes, 665 liver genes and 1057 lung genes. The significantly enriched pathways (adjusted p-val < 0.05) are presented in Table 5 and S3-S8 Tables (S1 File) (presenting the pathways (p-val < 0.05) with their p-values and corresponding marker genes). The "Cytokine-cytokine receptor interaction" and the "Viral protein interaction with cytokine and cytokine receptor" pathways were significantly enriched (adjusted p-value < 0.05) in all tissues. In addition, it can be seen from S3-S8 Tables (S1 File) that the "Amoebiasis" pathway is enriched (p-val < 0.05) in all six tissues. The "Asthma" pathway is enriched in subcutaneous adipose, liver, nerve tibial and skin. The "Staphylococcus aureus infection" pathway is enriched in subcutaneous adipose, lung, nerve tibial and skin.

Discussion

In this research we present a large-scale study testing the changes in gene expression and exploring marker genes of MV vs. non-MV samples from GTEx [22] donors, across six human tissues—lung, liver, muscle-skeletal, adipose-subcutaneous, skin and nerve-tibial. We developed 18 machine learning models, three models for each of the six tissues, using the XGBoost, RF and ANN methods to evaluate their predictive power for MV vs. non-MV samples and for feature analysis purposes. Our results show that the three methods can distinguish MV from non-MV samples successfully for these six tissues and that XGBoost outperforms the other

Table 5. Pathway enrichment analysis of the ventilation predictive marker genes across the six tested tissues.

	Tissue	Top pathways (adjusted p-value < 0.05)
1	Adipose Sub.	Viral protein interaction with cytokine and cytokine receptor
		Cytokine-cytokine receptor interaction
		Amoebiasis
		Chemokine signalling pathway
2	Liver	Cytokine-cytokine receptor interaction
		Viral protein interaction with cytokine and cytokine receptor
		Chemokine signalling pathway
		Neuroactive ligand-receptor interaction
3	Lung	Cytokine-cytokine receptor interaction
		Viral protein interaction with cytokine and cytokine receptor
		Neuroactive ligand-receptor interaction
		Chemokine signalling pathway
		Taste transduction
4	Muscle	Cytokine-cytokine receptor interaction
		Viral protein interaction with cytokine and cytokine receptor
		Chemokine signalling pathway
5	Nerve Tibial	Viral protein interaction with cytokine and cytokine receptor
		Cytokine-cytokine receptor interaction
		Neuroactive ligand-receptor interaction
6	Skin	Cytokine-cytokine receptor interaction
		Viral protein interaction with cytokine and cytokine receptor
		Chemokine signalling pathway
		Neuroactive ligand-receptor interaction

https://doi.org/10.1371/journal.pone.0264919.t005

methods, with an average accuracy of 0.951 and average AUC of 0.945 across the six tested tissues. The accuracy and AUC scores for the XGBoost models were higher than the ANN and RF models in all metrics aside from recall, in which the Neural Network model outperforms XGBoost but only by a small margin (all metrics > 0.93). Feature analysis showed that most significant genes affecting the prediction were related to tissue development, movement regulation, fibrosis and inflammation. We furthered explored marker gene convergence across tissues. Enrichment analysis of the marker genes showed significant enrichment of cytokine and viral signals for the tested tissues.

The importance of this research is not only in the ability to precisely predict the death circumstances based on the gene levels, but also in the analysis of the features that the machine learning model finds to be significant. Examining the different genes as part of a feature importance analysis reveals unexpected results, in the sense that we detect noteworthy genes or gene combinations with rather distinct value differences between the MV and non-MV groups. We detected tissue specific MV marker genes and marker genes shared across several tissues. Among the shared genes we detected CLEC3B, which is one of the most discriminant genes in both muscle and adipose (see Table 4) and is decreased in the MV group (see Fig 3A and S3 Fig in S1 File). The CLEC3B gene, which expresses the Tetranectin protein, has been reported to regulate muscle development [42] and is dysregulated in tumor tissues [44], which may explain its lower values in the MV group (see Fig 3A) which is assumed to have decreased muscle development following ventilation. The MXRA5 and EGR1 genes are both strong predictors of ventilation. MXRA5 is a highly significant (see Table 4) gene in the lungs, adipose-subcutaneous and skin-sun exposed tissues and EGR1 in muscle, nerve and adipose. Both the

MXRA5 and EGR1 genes were shown to be related to myocardial injury and dysregulated in on-pump vs. off-pump coronary artery bypass surgery [45]. In addition, MXRA5 is an apoptosis and remodeling marker that is associated with the ventilation injury-related response [45] and is a biomarker of severe respiratory syncytial virus (RSV) infection [46]. The tissue-specific angiotensin II receptor 2 (AGTR2) gene is expressed in lung fibrosis [43] and is shown by our analysis to be highly regulated in the lung's ventilation group (see Fig 3B). HRG is a high predictor of MV and is upregulated in the non-MV group (see Fig 3B). HRG gene levels are decreased in advanced lung cancer and the gene is known to have antifibrinolytic properties [47] that support its detected low levels in MV samples, which may indicate the progression of lung fibrosis. The TET1 gene, a significant predictor of MV in the muscle, is related to osteogenesis and adipogenesis inhibition [48]. The two top ranked liver marker genes are the TOR1A gene, which is related to smooth physical movements in the brain [49], and DNAJB4, a tumor suppressor gene [50]. In addition, we detected multiple carcinogenic and cancer marker genes among the top high predictor genes of ventilation such as VENTX [51], detected in adipose, MGMT [52] and DNAJB4 [50], detected in the liver, and GKN1 [53], which is an anti-inflammatory protein [54], detected in muscle and adipose with lower expression levels for the ventilation group.

Further enrichment analysis of the genes used for the predictions in each tissue indicated an inflammatory and viral gene signatures (see Table 5). Moreover, the "Amoebiasis", "Asthma" and "Staphylococcus aureus infection" pathways were enriched (p-val < 0.05, see S3-S8 Tables in S1 File) in multiple tissues including in the lungs. These pathways were associated with MV. For example, staphylococcus aureus is related to ventilator-associated pneumonia (VAP) [55] and amoeba-associated bacteria was suggested to be a cause of VAP in intensive care units [56]. In support of our results for the lungs, we mention McCall et al. [15] who, using the lung GTEx gene expression data, detected a large cluster of genes associated with type II pneumocytes related to cells that proliferate in ventilator associated lung injury. One explanation for the inflammatory and viral signatures we observed across tissues in our findings may be that ventilator-induced lung injury initiates non-pulmonary whole-body organ dysfunction [57].

As research limitations, we detected changes in gene expression between donors that were connected to MV prior to death and donors that were not. We note that these changes may be related to any direct and indirect factor related to the MV constellation, the MV machines, the patient's extended period of immobility (we indeed detected decreases in gene levels related to movement and development), and any other factor related to being under MV. In addition, since the GTEX tissues samples are derived from post-mortem but relatively healthy donors at time of death, we assume a gene distribution similarity between the deceased and living MV patients.

It is also worth noting that the GTEx tissue data includes bulk gene expression data composed of various cell types. Some gene expression changes we detected may be related to changes in the proportions in the cellular composition of the tissues. For example, the inflammatory signals may be related to changes in the proportions of immune-related cells in the tissues and not merely changes in the expression of genes within the cells. Nevertheless, these marker gene levels predicted MV and non-MV samples successfully and are a significant explanatory tool to understand ventilation induced changes either via gene expression changes within cells or proportion changes of cells in the tissues.

We note that we may further improve our performance by utilizing various extensions. For example, Reddy et al. [58] tested four popular machine learning methods and showed that if the dataset is of a high dimensionality, by performing a Principal Component Analysis (PCA) as a preprocessing step, it is possible to reach higher accuracy rates for the model. Nevertheless,

our focus is to detect a signature of single genes and not merely increase the performance. As further future work, we may use differential network analysis approach to gain further knowledge of the genes networks that changed between the MV and non-MV samples, e.g. Basha et al. [59] used differential network analysis to detect the changes in gene networks across human tissues.

In conclusion, we showed that MV induced transcriptomic changes in six tissues, going beyond the known direct effect on the lungs [21], and related to inflammation, bacterial and viral infections, fibrosis, tissue development, growth and movement regulation across all tested tissues. The changes in gene levels that we detected are highly significant and consistent across direct and peripheral tissues and thus MV should be carefully considered before being given to patients.

Supporting information

S1 File.
(PDF)

Acknowledgments

We thank Dr. Einat Minkov and Dr. Itay Dattner for the discussion and consultation.

Author Contributions

Conceptualization: Judith Somekh.

Formal analysis: Nir Lotan, Ehud Sussman, Gur Arye Yehuda.

Funding acquisition: Judith Somekh.

Investigation: Judith Somekh.

Methodology: Judith Somekh, Nir Lotan, Ehud Sussman.

Software: Ehud Sussman, Gur Arye Yehuda.

Supervision: Judith Somekh.

Validation: Nir Lotan.

Writing – original draft: Judith Somekh, Nir Lotan, Ehud Sussman, Gur Arye Yehuda.

Writing – review & editing: Judith Somekh.

References

1. Plani N., Becker P., and Van Aswegen H., "The use of a weaning and extubation protocol to facilitate effective weaning and extubation from mechanical ventilation in patients suffering from traumatic injuries: A non-randomized experimental trial comparing a prospective to retrospective cohort," *Physiother. Theory Pract.*, vol. 29, no. 3, pp. 211–221, Apr. 2013. https://doi.org/10.3109/09593985.2012.718410 PMID: 22943632
2. Gharib S. A., Liles W. C., Klaff L. S., and Altemeier W. A., "Noninjurious mechanical ventilation activates a proinflammatory transcriptional program in the lung," *Physiol. Genomics*, vol. 37, no. 3, pp. 239–248, May 2009. https://doi.org/10.1152/physiolgenomics.00027.2009 PMID: 19276240
3. Yu L., et al., "Machine learning methods to predict mechanical ventilation and mortality in patients with COVID-19," *PLoS One*, vol. 16, no. 4 April, p. e0249285, Apr. 2021. https://doi.org/10.1371/journal.pone.0249285 PMID: 33793600
4. Dolinay T., et al., "Gene expression profiling of target genes in ventilator-induced lung injury," https://doi.org/10.1152/physiolgenomics.00110.2005, vol. 26, no. 1, pp. 68–75, Jun. 2006.

5. Koenig S. M. and Truwit J. D., "Ventilator-associated pneumonia: Diagnosis, treatment, and prevention," *Clinical Microbiology Reviews*, vol. 19, no. 4. pp. 637–657, Oct-2006. https://doi.org/10.1128/CMR.00051-05 PMID: 17041138

6. Chastre J. and Fagon J. Y., "Ventilator-associated pneumonia," *American Journal of Respiratory and Critical Care Medicine*, vol. 165, no. 7. American Thoracic Society, pp. 867–903, 01-Apr-2002. https://doi.org/10.1164/ajrccm.165.7.2105078 PMID: 11934711

7. Tang H. and Shrager J. B., "The signaling network resulting in ventilator-induced diaphragm dysfunction," *American Journal of Respiratory Cell and Molecular Biology*, vol. 59, no. 4. American Thoracic Society, pp. 417–427, 01-Oct-2018. https://doi.org/10.1165/rcmb.2018-0022TR PMID: 29768017

8. Glau C. L., et al., "Progressive diaphragm atrophy in pediatric acute respiratory failure," *Pediatr. Crit. Care Med.*, vol. 19, no. 5, pp. 406–411, May 2018. https://doi.org/10.1097/PCC.0000000000001485 PMID: 29406380

9. Brower R. G., et al., "Higher versus Lower Positive End-Expiratory Pressures in Patients with the Acute Respiratory Distress Syndrome," *N. Engl. J. Med.*, vol. 351, no. 4, pp. 327–336, Jul. 2004. https://doi.org/10.1056/NEJMoa032193 PMID: 15269312

10. Esteban A., et al., "Characteristics and outcomes in adult patients receiving mechanical ventilation: A 28-day international study," *J. Am. Med. Assoc.*, vol. 287, no. 3, pp. 345–355, Jan. 2002. https://doi.org/10.1001/jama.287.3.345 PMID: 11790214

11. Hepokoski M. L., Malhotra A., Singh P., and Crotty Alexander L. E., "Ventilator-induced kidney injury: Are novel biomarkers the key to prevention?," *Nephron*, vol. 140, no. 2. S. Karger AG, pp. 90–93, 01-Sep-2018. https://doi.org/10.1159/000491557 PMID: 29996132

12. Bilotta F., Giordano G., Sergi P. G., and Pugliese F., "Harmful effects of mechanical ventilation on neurocognitive functions," *Critical Care*, vol. 23, no. 1. BioMed Central Ltd., 06-Aug-2019. https://doi.org/10.1186/s13054-019-2546-y PMID: 31387627

13. Pinhu L., Whitehead T., Evans T., and Griffiths M., "Ventilator-associated lung injury," *Lancet*, vol. 361, no. 9354. Elsevier B.V., pp. 332–340, 25-Jan-2003. https://doi.org/10.1016/S0140-6736(03)12329-X PMID: 12559881

14. Huang S., Kaipainen A., Strasser M., and Phd S. B., "Mechanical ventilation stimulates expression of the SARS-Cov-2 receptor ACE2 in the lung and may trigger a vicious cycle," May 2020.

15. McCall M. N. N., Illei P. B. B., and Halushka M. K. K., "Complex Sources of Variation in Tissue Expression Data: Analysis of the GTEx Lung Transcriptome," *Am. J. Hum. Genet.*, vol. 99, no. 3, pp. 624–635, Sep. 2016. https://doi.org/10.1016/j.ajhg.2016.07.007 PMID: 27588449

16. Gharib S. A., Liles W. C., Matute-Bello G., Glenny R. W., Martin T. R., and Altemeier W. A., "Computational identification of key biological modules and transcription factors in acute lung injury," *Am. J. Respir. Crit. Care Med.*, vol. 173, no. 6, pp. 653–658, Mar. 2006. https://doi.org/10.1164/rccm.200509-1473OC PMID: 16387799

17. Yen X. S., et al., "Interaction between regional lung volumes and ventilator-induced lung injury in the normal and endotoxemic lung," *Am. J. Physiol.—Lung Cell. Mol. Physiol.*, vol. 318, no. 3, pp. L494–L499, Mar. 2020. https://doi.org/10.1152/ajplung.00492.2019 PMID: 31940217

18. Liu J., Wu J., Liu S., Li M., Hu K., and Li K., "Predicting mortality of patients with acute kidney injury in the ICU using XGBoost model," *PLoS One*, vol. 16, no. 2 February, p. e0246306, Feb. 2021. https://doi.org/10.1371/journal.pone.0246306 PMID: 33539390

19. N. L. Loo, Y. S. Chiew, C. P. Tan, G. Arunachalam, A. M. Ralib, and M. B. Mat-Nor, "A MACHINE LEARNING MODEL FOR REAL-TIME ASYNCHRONOUS BREATHING MONITORING," in IFAC-*PapersOnLine*, 2018, vol. 51, no. 27, pp. 378–383.

20. Grunwell J. R., et al., "Machine Learning–Based Discovery of a Gene Expression Signature in Pediatric Acute Respiratory Distress Syndrome," *Crit. Care Explor.*, vol. 3, no. 6, p. e0431, Jun. 2021. https://doi.org/10.1097/CCE.0000000000000431 PMID: 34151274

21. Cai Y., Zhang W., Zhang R., Cui X., and Fang J., "Combined use of three machine learning modeling methods to develop a ten-gene signature for the diagnosis of ventilator-associated pneumonia," *Med. Sci. Monit.*, vol. 26, pp. e919035–1, Feb. 2020. https://doi.org/10.12659/MSM.919035 PMID: 32031163

22. Ardlie K. G., et al., "The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans," *Science (80-.).*, vol. 348, no. 6235, pp. 648–660, May 2015. https://doi.org/10.1126/science.1262110 PMID: 25954001

23. Ferreira P. G., et al., "The effects of death and post-mortem cold ischemia on human tissue transcriptomes," *Nat. Commun.*, vol. 9, no. 1, pp. 1–15, Dec. 2018. https://doi.org/10.1038/s41467-017-02088-w PMID: 29317637

24. F. T. Liu, K. M. Ting, and Z. H. Zhou, "Isolation forest," in *Proceedings—IEEE International Conference on Data Mining*, ICDM, 2008, pp. 413–422.
25. Somekh J., Shen-Orr S. S. S. S., and Kohane I. S. I. S., "Batch correction evaluation framework using a-priori gene-gene associations: Applied to the GTEx dataset," *BMC Bioinformatics*, vol. 20, no. 1, p. 268, May 2019. https://doi.org/10.1186/s12859-019-2855-9 PMID: 31138121
26. T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pp. 785–794, 2016.
27. Abiodun O. I., Jantan A., Omolara A. E., Dada K. V., Mohamed N. A. E., and Arshad H., "State-of-the-art in artificial neural network applications: A survey," *Heliyon*, vol. 4, no. 11, p. e00938, Nov. 2018. https://doi.org/10.1016/j.heliyon.2018.e00938 PMID: 30519653
28. Biau G. and Fr G. B., "Analysis of a Random Forests Model," *J. Mach. Learn. Res.*, vol. 13, pp. 1063–1095, 2012.
29. "scikit-learn: machine learning in Python—scikit-learn 1.0.2 documentation." [Online]. https://scikit-learn.org/stable/. [Accessed: 26-Feb-2022].
30. Labs T., Avenue P., Room A., Park F., and Schapire R. E., "A Brief Introduction to Boosting Analyzing the training error," *Proc. Sixt. Int. Jt. Conf. Artif. Intell.*, pp. 1401–1406, 1999.
31. Friedman J., "Greedy Function Approximation: A Gradient Boosting Machine," *Ann. Stat.*, vol. 29, no. 5, pp. 1189–1232, 2001.
32. Dimitrakopoulos G. N., Vrahatis A. G., Sgarbas K., and Plagianakos V., "Pathway analysis using xgboost classification in biomedical data," *ACM Int. Conf. Proceeding Ser.*, pp. 1–6, 2018.
33. Adeola Azeez Ogunleye and Wang Qing-Guo, "XGBoost Model for Chronic Kidney Disease Diagnosis," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, p. 1, 2019.
34. "XGBoost Python Library," 2020.
35. D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015—Conference Track Proceedings*, 2015.
36. Agarap A. F., "Deep Learning using Rectified Linear Units (ReLU)," Mar. 2018.
37. F. Chollet and others, "Keras," 2015. [Online]. https://github.com/fchollet/keras
38. Abadi M., et al., "TensorFlow: A system for large-scale machine learning," *Proc. 12th USENIX Symp. Oper. Syst. Des. Implementation, OSDI 2016*, pp. 265–283, 2016.
39. Pedregosa F, et al., "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, no. Nov 2014, pp. 2825–30, 2014.
40. Lundberg S. M. and Lee S. I., "A unified approach to interpreting model predictions," *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, no. Section 2, pp. 4766–4775, 2017.
41. Chen E. Y., et al., "Enrichr: Interactive and collaborative HTML5 gene list enrichment analysis tool," *BMC Bioinformatics*, vol. 14, Apr. 2013. https://doi.org/10.1186/1471-2105-14-128 PMID: 23586463
42. Wewer U. M., et al., "Tetranectin is a novel marker for myogenesis during embryonic development, muscle regeneration, and muscle cell differentiation in vitro," *Dev. Biol.*, vol. 200, no. 2, pp. 247–259, Aug. 1998. https://doi.org/10.1006/dbio.1998.8962 PMID: 9705231
43. Königshoff M., et al., "The angiotensin II receptor 2 is expressed and mediates angiotensin II signaling in lung fibrosis," *Am. J. Respir. Cell Mol. Biol.*, vol. 37, no. 6, pp. 640–650, Dec. 2007. https://doi.org/10.1165/rcmb.2006-0379TR PMID: 17630322
44. Liu J., et al., "CLEC3B is downregulated and inhibits proliferation in clear cell renal cell carcinoma," *Oncol. Rep.*, vol. 40, no. 4, pp. 2023–2035, Oct. 2018. https://doi.org/10.3892/or.2018.6590 PMID: 30066941
45. Ghorbel M. T., Cherif M., Mokhtari A., Bruno V. D., Caputo M., and Angelini G. D., "Off-pump coronary artery bypass surgery is associated with fewer gene expression changes in the human myocardium in comparison with on-pump surgery," *Physiol. Genomics*, vol. 42, no. 1, pp. 67–75, Jun. 2010. https://doi.org/10.1152/physiolgenomics.00174.2009 PMID: 20332183
46. Van Den Kieboom C. H., et al., "Nasopharyngeal gene expression, a novel approach to study the course of respiratory syncytial virus infection," *Eur. Respir. J.*, vol. 45, no. 3, pp. 718–725, Mar. 2015. https://doi.org/10.1183/09031936.00085614 PMID: 25261323
47. Winiarska A., Zareba L., Krolczyk G., Czyzewicz G., Zabczyk M., and Undas A., "Decreased levels of histidine-rich glycoprotein in advanced lung cancer: Association with prothrombotic alterations," *Dis. Markers*, vol. 2019, 2019. https://doi.org/10.1155/2019/8170759 PMID: 30944671
48. Cakouros D., et al., "Specific functions of TET1 and TET2 in regulating mesenchymal cell lineage determination 06 Biological Sciences 0601 Biochemistry and Cell Biology 06 Biological Sciences 0604

Genetics 11 Medical and Health Sciences 1103 Clinical Sciences," *Epigenetics and Chromatin*, vol. 12, no. 1, p. 3, Jan. 2019. https://doi.org/10.1186/s13072-018-0247-4 PMID: 30606231

49. "TOR1A gene: MedlinePlus Genetics." [Online]. https://medlineplus.gov/genetics/gene/tor1a/#resources. [Accessed: 10-Feb-2021].

50. Acun T., et al., "HLJ1 (DNAJB4) Gene Is a Novel Biomarker Candidate in Breast Cancer," *Omi. A J. Integr. Biol.*, vol. 21, no. 5, pp. 257–265, May 2017. https://doi.org/10.1089/omi.2017.0016 PMID: 28481734

51. Rawat V. P. S., et al., "The vent-like homeobox gene VENTX promotes human myeloid differentiation and is highly expressed in acute myeloid leukemia," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 107, no. 39, pp. 16946–16951, Sep. 2010. https://doi.org/10.1073/pnas.1001878107 PMID: 20833819

52. Gerson S. L., "MGMT: Its role in cancer aetiology and cancer therapeutics," *Nature Reviews Cancer*, vol. 4, no. 4. Nature Publishing Group, pp. 296–307, 2004. https://doi.org/10.1038/nrc1319 PMID: 15057289

53. Dokhaee F., Mazhari S., Galehdari M., Bahadori Monfared A., and Baghaei K., "Evaluation of GKN1 and GKN2 gene expression as a biomarker of gastric cancer.," *Gastroenterol. Hepatol. from bed to bench*, vol. 11, no. Suppl 1, pp. S140–S145, 2018. PMID: 30774821

54. Chin S. C. N., et al., "Coordinate expression loss of GKN1 and GKN2 in gastric cancer via impairment of a glucocorticoid-responsive enhancer," *Am. J. Physiol.—Gastrointest. Liver Physiol.*, vol. 319, no. 2, pp. G175–G188, Aug. 2020. https://doi.org/10.1152/ajpgi.00019.2020 PMID: 32538140

55. Burnham J. P. and Kollef M. H., "Prevention of Staphylococcus aureus Ventilator-Associated Pneumonia: Conventional Antibiotics Won't Cut It," *Clin. Infect. Dis. An Off. Publ. Infect. Dis. Soc. Am.*, vol. 64, no. 8, p. 1089, Apr. 2017.

56. La Scola B., Boyadjiev I., Greub G., Khamis A., Martin C., and Raoult D., "Amoeba-Resisting Bacteria and Ventilator-Associated Pneumonia," *Emerg. Infect. Dis.*, vol. 9, no. 7, p. 815, Jul. 2003. https://doi.org/10.3201/eid0907.020760 PMID: 12890321

57. Brander L. and Slutsky A. S., "Does ventilator-induced lung injury initiate non-pulmonary organ dysfunction?," in *Intensive Care Medicine*: *Annual Update 2006*, Springer New York, 2007, pp. 424–434.

58. Reddy G. T., et al., "Analysis of Dimensionality Reduction Techniques on Big Data," *IEEE Access*, vol. 8, pp. 54776–54788, 2020.

59. Basha O., et al., "Differential network analysis of multiple human tissue interactomes highlights tissue-selective processes and genetic disorder genes," *Bioinformatics*, vol. 36, no. 9, pp. 2821–2828, May 2020. https://doi.org/10.1093/bioinformatics/btaa034 PMID: 31960892