healthcare

MDPI

Article

Machine Learning Model Based on Lipidomic Profile Information to Predict Sudden Infant Death Syndrome

Karen E. Villagrana-Bañuelos, Carlos E. Galván-Tejada *, Jorge I. Galván-Tejada, Hamurabi Gamboa-Rosales, José M. Celaya-Padilla, Manuel A. Soto-Murillo and Roberto Solís-Robles

Unidad Académica de Ingeniería Eléctrica, Universidad Autónoma de Zacatecas, Jardín Juárez 147, Centro, Zacatecas 98000, México; kvillagrana@uaz.edu.mx (K.E.V.-B.); gatejo@uaz.edu.mx (J.I.G.-T.); hamurabigr@uaz.edu.mx (H.G.-R.); jose.celaya@uaz.edu.mx (J.M.C.-P.); alejandro.somu@uaz.edu.mx (M.A.S.-M.); rsolis@uaz.edu.mx (R.S.-R.)

* Correspondence: ericgalvan@uaz.edu.mx

Abstract: Sudden infant death syndrome (SIDS) represents the leading cause of death in under one year of age in developing countries. Even in our century, its etiology is not clear, and there is no biomarker that is discriminative enough to predict the risk of suffering from it. Therefore, in this work, taking a public dataset on the lipidomic profile of babies who died from this syndrome compared to a control group, a univariate analysis was performed using the Mann–Whitney *U* test, with the aim of identifying the characteristics that enable discriminating between both groups. Those characteristics with a *p*-value less than or equal to 0.05 were taken; once these characteristics were obtained, classification models were implemented (random forests (RF), logistic regression (LR), support vector machine (SVM) and naive Bayes (NB)). We used seventy percent of the data for model training, subjecting it to a cross-validation (k = 5) and later submitting to validation in a blind test with 30% of the remaining data, which allows simulating the scenario in real life—that is, with an unknown population for the model. The model with the best performance was RF, since in the blind test, it obtained an AUC of 0.9, specificity of 1, and sensitivity of 0.8. The proposed model provides the basis for the construction of a SIDS risk prediction computer tool, which will contribute to prevention, and proposes lines of research to deal with this pathology.

Keywords: SIDS; lipidomic; metabolomic; glycerophospholipids; machine learning; biomarker

Citation: Villagrana-Bañuelos, K.E.; Galván-Tejada, C.E; Galván-Tejada, J.I.; Gamboa-Rosales, H.; Celaya-Padilla, J.M.; Soto-Murrillo, M.A.; Solís-Robles, R. Machine Learning Model Based on Lipidomic Profile Information to Predict Sudden Infant Death Syndrome. *Healthcare* **2022**, *10*, 1303. https://doi.org/10.3390/healthcare10071303

Academic Editor: Tin-Chih Toly Chen

Received: 12 May 2022
Accepted: 9 July 2022
Published: 14 July 2022

1. Introduction

Sudden infant death syndrome (SIDS) is the leading cause of infant mortality after the neonatal period in developed countries [1], and it is defined as death that remains unexplained, after having exhausted clinical and forensic investigations in children under one year of age [2]. Therefore, it is considered a diagnosis of exclusion; there are multiple theories that try to explain what causes this death without fully clarifying the physiopathogenesis of how or why it occurs. One of the most accepted theories is the triple risk theory, which implies that a baby is vulnerable, in a critical period of development, and there was a trigger. Where there are genetic factors that alter the central nervous system, cardiac channelopathies have been associated [3,4]; as well as inborn errors of metabolism; brainstem dysfunction and cardiorespiratory function; respiratory obstruction; infections; heat stress; and compression of vertebral arteries; among many other factors [5]. Taking this and modifiable risk factors into account, attempts to avoid these deaths have focused on prevention, especially safe sleep [1]; however, tools need to be investigated to help identify these patients. Despite the success obtained from the safe sleep campaigns, inquiry is required in order to obtain answers, which is why, at the moment, the realization of postmortem directed genetic tests (molecular autopsy) is recommended within the investigation of the cases [3].

Lipidomics and metabolomics have been described in the literature as optimal approaches to find differences in a particular physiological state. For example, a study of the metabolic profile in urine of preterm newborns of mothers with and without chorioamnioitis was carried out, where they found differences between the metabolites that were able to distinguish those born to mothers who suffered from this pathology or not [6]. Other authors [7] have worked with the metabolic profile in postmortem brains and identified changes in the urea as well as changes in carbohydrate and protein metabolism.

Metabolomics currently in disease research shows promise for biomarker discovery. Biomarkers are useful for predicting or identifying a risk of disease, diagnosing it, monitoring or prognosis [8]. Lipidomics is found within the metabolomics [9]; independently, it is dedicated to the study and characterization of the set of cellular lipids, the molecules with which they interact and their functions in the body, showing that the classic functions with which they are associated with lipids are those of structure and energy storage. However, technological advances have shown that there are different lipids in the human body suggesting the existence of functions not yet explored [10]. For example, in the detection of unexplored pathophysiological mechanisms for brain diseases [11], the potential value of metabolomics to study SIDS stands out. This study compares the profiles of the medulla oblongata in human brains of patients who died from SIDS with a control group, concluding that this type of study could lead to the antemortem identification of biomarkers [12]. Graham et al. directed a study of brain metabolomics in patients who died from SIDS compared to a control group, and they were able to identify possible biomarkers [13].

Omics technologies provide a large amount of data, which, when used in a univariate or multivariate manner, could contribute to the detection of the risk profile of the patient who dies from SIDS [14]. The large amount of data, which is produced with new technologies, benefits from the analysis with machine learning techniques. In the literature, it is described that through using machine learning (ML) techniques with characteristics between groups of cases and controls, it is possible to find biomarkers that help in medical diagnosis [15–18].

Decision making in medical diagnosis is a complicated process in which various factors are involved that can affect the certainty of doctors; for this reason, different ML techniques have been implemented to improve certainty in the diagnosis of diseases [19]. For example, Zoabi and collaborators used this type of technique in conjunction with clinical data to predict coronavirus disease [20]. Other researchers have implemented ML models for the prediction of cardiac arrhythmias [21]. ML has also been applied to cancer diagnosis [22], and it has also been recently applied to SIDS. For example, Blackburn and collaborators, taking into account infant mortality files and using unsupervised machine learning techniques, have identified groups of descendants of SIDS [23]. Galván-Tejada and collaborators; using clinical and short-chain fatty acid data, have also sought to aid in the diagnosis of SIDS [24]. However, it is necessary to continue investigating to identify the real cause of SIDS and effective models, which allow reducing more deaths from this pathology.

For this reason, in this work, the lipidomics of patients who died from SIDS are analyzed in comparison with control patients, and machine learning classification models are implemented in order to contribute to finding the risk profile and potential antemortem biomarkers that allow avoiding more deaths in the future.

2. Materials and Methods

In this section, the materials and methods used are briefly described, and the steps that were carried out in the experimentation stage are visualized in Figure 1. All of the experimentation was developed in R (version 4.1.0) [25], which is a free software environment for graphics and statistical computing.

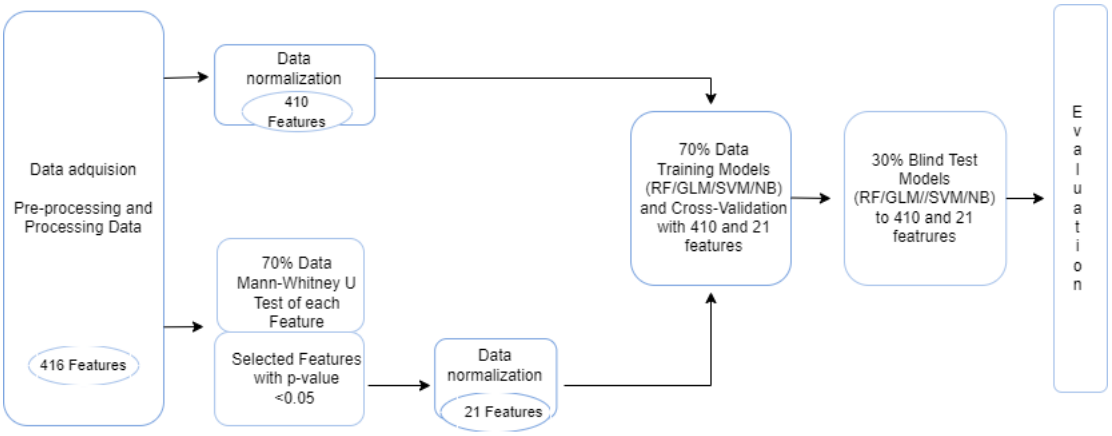

Figure 1. Flowchart of the steps that were followed for the development of the machine learning models, until their evaluation.

A public dataset available in the National Metabolomics Data Repository of the NIH Common Fund was used for this study. The dataset is named Lipidomics in sudden infant death syndrome, and it takes into account deceased patients diagnosed with SIDS (cases) and deceased patients from any other cause (controls) [26]. Lipid values were extracted from serum samples, and everything related to their processing regarding clinical analysis can be found in the "sample preparation" section where the public dataset was found.

2.1. Data Description

The information contained in the dataset "Lipidomics in sudden infant death syndrome" includes 6 characteristics of clinical information (post-conception age, postnatal age, gestational age, patient identification numbers, and the class or diagnosis—that is, case or control patient. The first includes patients who died and were diagnosed as due to SIDS and the second corresponds to the death of the patient from any other cause other than SIDS). The rest of the information is divided into two groups; the first corresponds to the negative mode analysis of the C18 ion, in which there are 132 characteristics (lipids), and there are 278 in the positive C18 ion. In total, there are 416 characteristics obtained postmortem in 33 patients, of which 23 correspond to cases and the rest are controls.

Each of them was grouped by the super class; for easy viewing, Table 1 shows the first column, which corresponds to the group, and the second column corresponds to the number of features in this group for the ion C18 negative mode analysis. Table 2 shows the same distribution of information, but it corresponds to the ion C18 positive mode analysis. So, joining both modes of analysis, we have 16 different groups of lipids.

Table 1. Grouped features according to their super class; ION C18 negative mode analysis.

Group	Number of Features
Cardiolipins	6
Sphingolipids	1
Acids	16
Glycerophosphate	1
Phosphatylcholine	24
Phosphatylethalonamine	24
Phosphatidylglycerols	15
Phosphatidylinositols	12
Glycerophosphoserines	9
Lysophosphatidylethanolamine	8
Ether Phosphatidylethanolamines	16

Table 2. Grouped features according to their super class; ION C18 positive mode analysis.

Group	Number of Features
Cholesterol esters	12
Diacylglycerols	37
Monoradylglycerols	2
Phosphatylcholine	37
Phosphatylethalonamine	11
Sphingomyelins	43
Triacylglycerols	98
Lysophosphatidylcholines	25
Ether Phosphatidylethanolamines	4
Ether Phosphatidylcholines	9

2.2. Data Preprocessing

In this work, only the characteristics corresponding to lipids were included, uniting in a single data set both the mode analysis positive and negative C18 ion characteristics. So, 410 characteristics were used in total without including the class or diagnostic feature. We included 33 patients; 23 of them were classified as death by SIDS and the rest died from any other diagnosis, so they are considered control cases.

An exhaustive search was carried out to look for missing data, which were not found; therefore, there was no need to impute the data. The diagnostic variable was converted into binary values, representing 0 for control patients and 1 for SIDS cases.

2.3. Data Normalization

For the normalization of the data, the conversion of the values to *z*-scores was carried out, which represents how many standard deviations below or above the mean is the value to be evaluated from a reference population. This is exemplified in Equation (1), where x is the observed measure, mu is the population mean, and $sigma$ is the population standard deviation [27].

$$z = \frac{x - mu}{\sigma} \tag{1}$$

2.4. Classification Methods

Classification is one of the uses of supervised learning, which aims to predict class labels. In the case of medicine, with a binary classification, it seeks to identify the diagnosis: whether a patient is sick or healthy. Each of the implemented methods is described below and were developed in R [25], a free software environment for statistical computing and graphics.

Random forest (RF) [28] is based on multiple decision trees, so it has the advantage of decreasing the possibility of overfitting; each tree has a prediction of class, and finally, there is a mean to determine to which class each patient belongs: for example, if a patient is sick or not.

Each tree follows a series of logical decisions, similar to a flow chart, with decision nodes indicating a decision to be made about an attribute. Each branch indicates the decision options, which are assigned a predicted class. An important decision is to choose the best split, so that the ideal is that the partitions contain examples of a single class. If this happens, they are considered pure segments. One of the measures to measure this purity is entropy; the minimum value of 0 indicates that the sample is completely homogeneous, while 1 indicates the maximum amount of disorder. In the entropy Equation (2), for a given segment of data (S), the term n refers to the number of different class levels, and P_i refers to the proportion of values falling into class level i. In simple terms, the total entropy resulting from a division is the sum of the entropy; each of the n partitions is weighted by the proportion of examples that fall in that partition (w_i) [29]. We used the package

'randomForest' [30]. It is important to point out that all the functions mentioned in this work were used with the pre-established hyperparameters.

$$Entropy(S) = \sum_{i=1}^{n} w_i Entropy(P_i) \tag{2}$$

Logistic regression (LR) [31] is used for binary classification, modeling the probability of belonging to one group or another. It transforms the value obtained with the linear regression ($\beta + \beta_1 X$), using a function; the most used is the sigmoid function (Equation (3)), which will result in a value between 0 and 1. In other words, it is useful when you are interested in the impact of different explanatory variables on a binary response variable [32]. We used the funtion '*glm*' into the package 'stats' [33].

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{3}$$

Support vector machine (SVM) is also useful for classification; it is considered a robust method, since it combines characteristics of the nearest neighbors and regression methods [29]. Its objective is to predict the new sample class, with the prior learning of the training. SVM uses a linear boundary called a hyperplane to divide data into groups of similar elements shown in Equation (4). Generally, as indicated by the class values, assuming you have two classes, the hyperplane, constructed, optimally separates, or whichever has the maximum distance (margin); a good separation between the classes will allow a correct classification [34].

$$\bar{w} \cdot \bar{x} + b = 0 \tag{4}$$

However, in practice, some relationships between variables are not linear, so a kernel function can be applied to transform the features, assigning the data to a different dimensional space to achieve separation between classes. There are different types of kernels, such as linear, polynomial, or sigmoid; on this occasion, the radial kernel was used, also known as the radial basis function (RBF), which is popular for its similarity to the Gaussian distribution. It is represented by Equation (5) [35]. The 'e1071' package was used with the 'svm' function [36].

$$K(x_1, x_2) = exp\left(-\frac{\|x_1 - x_2\|^2}{2\sigma^2}\right) \tag{5}$$

Naive Bayes (NB) is a method that searches to describe the probability of occurrence of an event. Taking into account the Bayes theorem [37], the probability of occurrence of an event A is sought; since event B has already occurred, it is also known as conditional probability, and it is exemplified in the following Equation (6). Thus, the probability is determined of a sample belonging to one class or another.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(A \bigcap B)}{P(B)} \tag{6}$$

The naive Bayes algorithm makes two assumptions about the data: the first is that the data are equally important and that they are independent, which is unlikely to be the case in reality. So, the general formula would be:

$$P(y_i \mid X_1, X_2..., X_n) = P(X_1 \mid y_i)P(X_2 \mid y_i)...P(X_n \mid y_i) \tag{7}$$

Such assumptions have lent themselves to error speculation; however, it is said to work well, since it is not important to obtain a careful probability estimate provided that the predicted class values are true. For instance, if a disease classification model correctly identifies sick a patient, it does not matter whether it was 51% or 99% as long as the predicted class is correct [29]. The 'naivebayes' function was used [36].

2.5. Feature Selection

The selection of the characteristics was carried out through a filtering method, which consists of applying different statistical techniques. In this case, considering that the sample size is small, according to the literature, the use of non-parametric tests is useful, which are used when the data do not present a Gaussian distribution. According to the literature, the Mann–Whitney *U* test [38] is useful when the normal or Gaussian distribution of the data cannot be identified, due to a small sample of the population, as is the case here. It takes into account the medians to know if there are differences between two independent populations [39], as in the case of patients who died from SIDS and control cases, according to each of the characteristics or lipids in this case.

The characteristics are ordered independently of the population to which they belong in ascending order to obtain the ranks. Subsequently, the Mann–Whitney *U* is calculated for each characteristic with Equations (8) and (9), where n_1 corresponds to the first sample population and n_2 corresponds to the second population, R represents the sum of the ranges of the population sample, respectively, obtaining the p value for 95% of statistical significance. The 'wilcox.test' function was used [40].

$$U_1 = n_1 n_2 + \frac{n_1(n_{1+1})}{2} - R_1 \tag{8}$$

$$U_2 = n_1 n_2 + \frac{n_2(n_{2+1})}{2} - R_2 \tag{9}$$

2.6. Cross-Validation

In this work, cross-validation was used, which is a statistical method that helps evaluate and compare learning algorithms. The main idea when using this method is that a part of the data set is hidden from the training model (the part shaded in orange in the Figure 2) in each iteration (k), and this repeats until at some point each of the folds created is used as a test. Then, we subject the model to a blind evaluation with the data that were kept out (the part shaded in purple in the Figure 2). This ensures that by subjecting the classification model to a new population, the algorithm will behave similarly to the test stage in the experimental phase, largely avoiding overfitting [41]. It was implemented with the function 'trainControl' inside the 'caret' package [42].

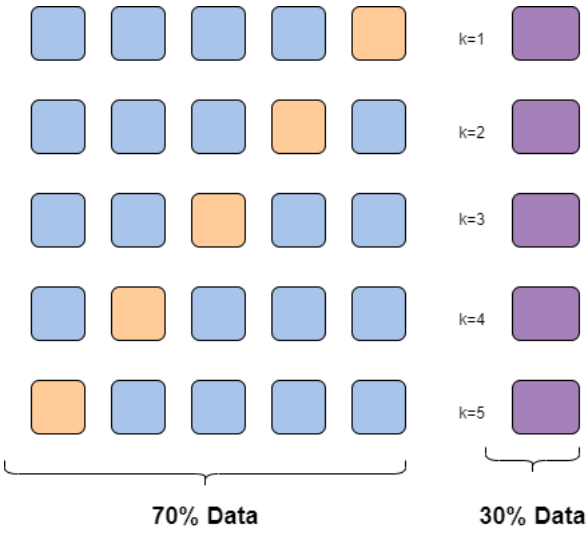

Figure 2. Exemplification of the cross-validation used in this work.

2.7. Metrics of Evaluation

There are two classification possibilities in this work: the SIDS patient or control case. Therefore, with the cases subjected to the model, a confusion matrix can be formed.

To determine the true positive rate (*Sensitivity*), that is, the probability that a sick subject has a positive test result, we use Equation (10), where *TP* refers to true positive cases (patients with disease classified as sick) and *FN* refers to false negatives (healthy patients classified as sick). The true negative rate (*Specificity*), in other words, is the probability that a healthy subject will have a negative test result as represented by Equation (11), where *TN* are true negatives (healthy patients classified as healthy) and *FP* are false positives (sick patients classified as healthy); based on this, the ROC curve (receiver operation curve) is plotted, which represents the probability of the classifier to correctly determine a sample chosen at random, either positive or negative; and the area under the curve (AUC) was calculated. The higher the AUC, the better the model is at predicting 0 as 0 (healthy) and 1 as 1 (sick). This means that the higher the AUC, the better the model will be at distinguishing between patients with disease and without disease. That is, the model is excellent if it has an AUC of 1, which is interpreted as that the model has a good measure of separability. On the other hand, if the AUC is close to 0, it is the worst measure of separability: that is to say that instead of classifying 1 as 1, it does so as 0, and vice versa. The AUC of 0.5 represents that the model has no class separation capability, which is equivalent to making a random decision [43–46].

$$Sensitivity = \frac{TP}{TP + FN} \tag{10}$$

$$Specificity = \frac{TN}{TN + FP} \tag{11}$$

Another metric used was *accuracy*, which is a statistical measure of how well a binary classification test correctly identifies or excludes a condition. In other words, it is the proportion of true results among the total number of cases examined and is represented by the following Equation (12), where *TP* are true positives (sick patients identified as such), *TN* are true negatives (healthy patients identified as such), *FP* are false positives (healthy patients identified as sick) and *FN* are false negatives (sick patients identified as healthy) [47]. For this stage, the 'caret' [48] and 'pROC' [49] libraries were used.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{12}$$

3. Results and Experimentation

This section presents the experiments carried out for the development of this work as well as the results obtained. As the first step, the set of lipidomic features (410 plus the class) were subjected to four different classification methods such as random forest (RF), logistic regression (RL), support vector machine (SVM), and naive Bayes (NB). These were used to predict 0 (control patient) or 1 (patient died from SIDS), using 70% of the data and cross-validation with *k*-folds (5). Subsequently, each of the trained models was submitted to evaluation with the remaining 30% of data, performing a blind test, and the evaluation metrics reported in Table 3 were obtained. Different rates for splitting the dataset are reported in the literature for the training and testing stages; however, it was observed that using the 70/30 rate on a small dataset ensures that the same proportion of cases and controls is maintained.

Table 3. Evaluation metrics for each classifier of the standardized dataset.

Classification Method	Features	AUC	Accuracy	Sensitivity	Specificity
RF	410	0.2857	0.4444	0.5714	0
RF	21	**0.9000**	**0.8889**	0.8000	**1**
LR	410	0.4500	0.5555	0.4000	0.7500
LR	21	0.7500	0.7777	0.8000	0.7500
SVM	410	0.7000	0.7777	0.8000	0.7500
SVM	21	**0.9000**	0.8888	**1**	0.7500
NB	410	0.6750	0.6666	0.6000	0.7500
NB	21	0.8000	0.7777	0.6000	**1**

Figure 3 shows the AUC obtained in the blind test of the models implemented with the 21 selected features. It is observed that regardless of the classification method used, all exceed 0.5 AUC, with the RF and SVM method achieving the best performance with a 0.9 AUC.

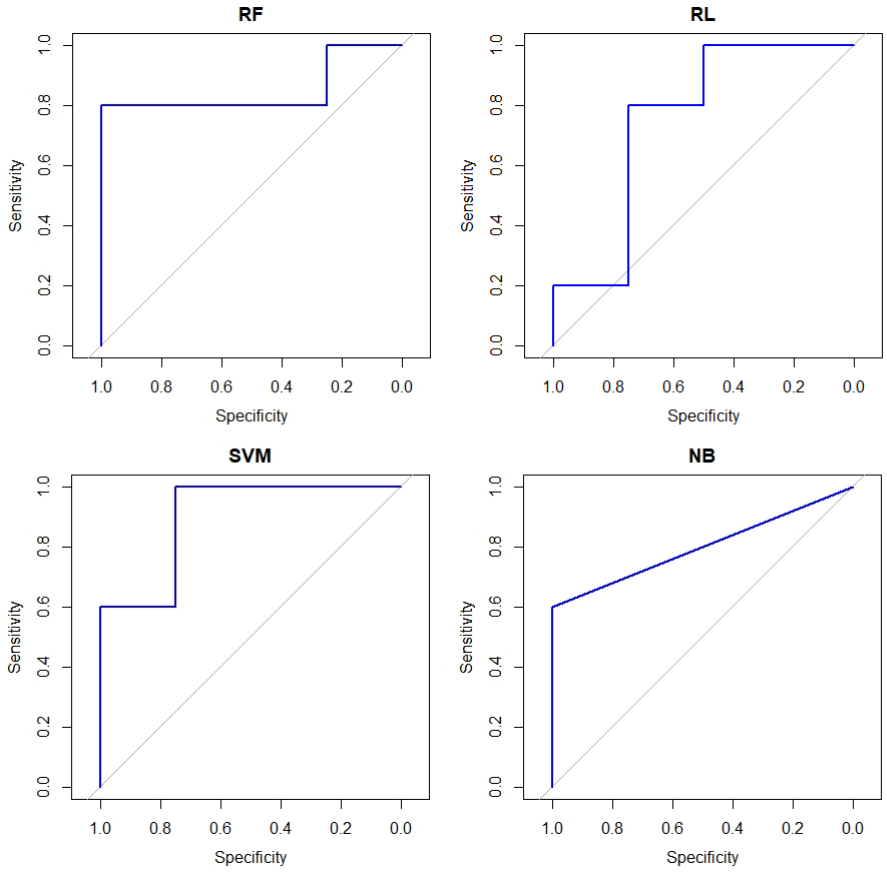

Figure 3. ROC curves of the four machine learning classification models, with the 21 selected features, blind test results.

Then, a new experiment was carried out, using 70% of the data, which consisted of submitting each of the 410 characteristics to the Mann–Whitney *U* test, with which the *p*-value was obtained. Once this analysis was carried out, the characteristics with the *p*-value greater than 0.05 were selected, with a total of 21 characteristics, which are shown in Table 4. In this table, the name of the characteristic is shown in the first column, while the corresponding class to which it belongs, the chemical formula and finally the *p*-value obtained are observed on the right.

Table 4. Features selected by Mann–Whitney *U* test of the lipidomic profile in SIDS.

Features	Super Class	Main Class	Sub Class [1]	Formula	*p*-Value
PC 40:7	Glycerophospholipids	Glycerophosphocholines	PC	$C_{48}H_{82}NO_8P$	0.00420
PI 36:2	Glycerophospholipids	Glycerophosphocholines	PC	$C_{44}H_{84}NO_8P$	0.00420
PE 35:0	Glycerophospholipids	Glycerophosphoethanolamines	PE	$C_{40}H_{80}NO_8P$	0.01060
DG 34:1	Glycerolipids	Diradylglycerols	DAG	$C_{37}H_{70}O_5$	0.01308
PC.38.7	Glycerophospholipids	Glycerophosphocholines	PC	$C_{46}H_{78}NO_8P$	0.01602
PE 34:3	Glycerophospholipids	Glycerophosphoethanolamines	PE	$C_{39}H_{72}NO_8P$	0.02355
TG 57:8	Glycerolipids	Triradylglycerols	TAG	$C_{60}H_{100}O_6$	0.02355
CL 70:5	Glycerophospholipids	Cardiolipins	CL	$C_{79}H_{144}O_{17}P_2$	0.02355
SM 40:1	Sphingolipids	Sphingomyelins	SM	$C_{45}H_{91}N_2O_6P$	0.02826
PC 30:2	Glycerophospholipids	Glycerophosphocholines	PC	$C_{38}H_{72}NO_8P$	0.02826
PC 32:3	Glycerophospholipids	Phosphatidylcholines	PC	$C_{40}H_{74}NO_8P$	0.03372
SM 36:2	Sphingolipids	Sphingomyelins	SM	$C_{41}H_{81}N_2O_6P$	0.03372
PC 33:1	Glycerophospholipids	Glycerophosphocholines	PC	$C_{41}H_{80}NO_8P$	0.03372
CE 18:2.	Sterol Lipids	Sterol esters	Chol	$C_{45}H_{76}O_2$	0.03372
DG 36:2	Glycerolipids	Diradylglycerols	DAG	$C_{39}H_{72}O_5$	0.03372
PC 32:1	Glycerophospholipids	Glycerophosphocholines	PC	$C_{38}H_{73}O_{10}P$	0.03999
PG 36:3	Glycerophospholipids	Glycerophosphoglycerols	PG	$C_{42}H_{77}O_{10}P$	0.04717
CE 22:6	Sterol Lipids	Sterol esters	Chol	$C_{49}H_{76}O_2$	0.04717
PC 40:10	Glycerophospholipids	Glycerophosphocholines	PC	$C_{50}H_{80}NO_8P$	0.04717
PC 42:7	Glycerophospholipids	Glycerophosphocholines	PC	$C_{50}H_{86}NO_8P$	0.04717
SM.30.1	Sphingolipids	Sphingomyelins	SM	$C_{35}H_{71}N_2O_6P$	0.04717

[1] PC—Phosphatidylcholines; PE—Phosphatidylethanolamines; DAG—Diacylglycerols; TAG—Triacylglycerols; CL—Cardiolipins; SM—Sphingomyelins; Chol—Cholesterol esters; PG—Phosphatidylglycerols.

Each of the selected characteristics was subjected to normalization, using a *z*-score. With these selected characteristics plus the class or diagnosis, classification models were trained with the RF, RL, SVM and NB methods, dividing the data by 70 and 30 for cross-validation and blind testing, respectively. The same were evaluated in each stage, and the performance of each of the models is reported in Table 3.

4. Discussion

These results represent the first report describing a complete lipid profile analysis of cases and controls in SIDS.

A first experiment was carried out, which consisted of training the classification models, which will identify patients at risk of dying from SIDS compared to the control group (who died for any other reason), using four machine learning methods (RF, RL, SVM and NB), using 410 features (lipids values).

In the second experiment, using the statistical method of the Mann–Whitney *U* test, it was sought to know if it was possible to differentiate between the two groups with these features. We found that only 21 characteristics achieved this objective with a *p*-value of <0.05, that is, achieving statistical significance. In addition, each of the selected characteristics was subjected to normalization by means of *z*-score, and four classification models were subsequently implemented with the same automatic learning methods of the previous experiment in order to carry out the comparison.

Both experiments were subjected to cross-validation and exposed to evaluation with 30% of the data to perform a blind test, which means that the developed models had no prior knowledge of the new data to be evaluated, obtaining evaluation metrics for each one of them. In Table 3, the blind test stage of each one of the models is observed, with its two variants, with all the characteristics (410) and with the selected ones (21). It is clear that the models with 21 features present better performance compared to the models with 410 features. This is consistent with the literature [50,51], since a greater number of features does not ensure better performance of machine learning models, but on the contrary, they can be detrimental, as in this case.

According to the results, in the test stage, any of the four models (with 21 characteristics) present acceptable evaluations with an AUC greater than 0.75, specificity greater than 0.75, sensitivity greater than 0.6 and accuracy of 0.7777. However, the RF method stands

out for having three of the four metrics evaluated with the best performance (AUC 0.9), which is followed by SVM (AUC 0.9), NB (AUC 0.8) and finally LR (AUC 0.75).

The NB model stands out by maintaining the same value in the accuracy and sensitivity metrics of the model with all and with the selected features. Based on the principle of independence of the characteristics for which this classifier is considered naive, it is difficult for such a statement to be true, since in the human body, metabolites and other body substances are in constant interaction. This model, according to the sensitivity, is the one with the worst performance regarding detecting the disease in sick subjects: it barely manages to overcome the random choice.

Such performances could be improved by balancing the data set. Different techniques allow for balancing data sets. For example, there are the resampling methods; within them are those of oversampling, which are responsible for increasing the minority class; on the other hand, those of subsampling reduce the minority class. However, they are controversial, since the use of subsampling techniques could lose valuable information, while oversampling techniques if the samples are few can lead to overfitting. For this reason, in this work, we decided not to implement them; however, it is interesting to know the behavior of the models when implementing these techniques and discuss them, which could be future work. Regarding the values close to the unit, as the sensitivity of 1, in the SVM model with 21 features, even using cross-validation techniques and the blind test would also benefit from testing the models with oversampling techniques or bigger data sets.

Of the lipids capable of discriminating between the two groups, 13 of the them correspond to the group of glycerophospholipids, of which 8 correspond to the group of glycerophosphocholines, 2 correspond to the group of glyphosphoethanolamines, 1 corresponds to the group of cardiolipins, 1 corresponds to the group of phosphatidylcholine, and 1 corresponds to the group of glycerophosphoglycerol. Of the remaining lipids, three belong to the group of sphingolipids and a subgroup of sphingomyelins, three belong to glycerolipids in subgroups of one triradyglyceroles and two diradylglycerols; and the remaining two belong to the sterol lipid group and sterol ester subgroup.

Hishikawa et al. [52] describe in their research the diversity of the functions of membrane glycerophospholipids in mammalian cells, which can be the constituents of cell membranes as well as precursors of cell signaling molecules and are part of lipoproteins and bile, among others. On the other hand, cone-shaped glycephospholipids with a small polar head (such as PE and CL), and/or bulky acyl chains, have important functions in membrane fusion, endocytosis, exocytosis, cytokinesis and in vesicle trafficking. That is to say that they are important in the transport of lipids. In addition, CL has been associated with problems of cellular respiration and energy production. It also stands out that lung surfactant, which is produced by type II pneumocytes, prevents lung collapse, and it is approximately 90% made up of lipids, and there are mainly dipalmitoyl-phosphatidylcholine and associated proteins in the remaining 10%. In other research, the role of glycerophospholipids in neural membranes is recognized, which is also relevant, since brain problems have been associated as causes of death from this syndrome [53–55]. This reinforces theories about pulmonary defects and at the same time opens possible lines of research since, when found in the results of this work, the ability to differentiate patients who died from SIDS from the control group, with glycerophospholipids, could lead to finding answers about the real origin of this pathology.

In the related literature, the work of Graham et al. [12,13] stands out, as they have used undirected metabolomics in their first work and directed metabolomics in their second to be able to identify babies who die of SIDS against those who do not, achieving AUC values of 1 and 0.92, respectively. However, they recognize some limitations in their work, such as that the cohort is not balanced, and that it is necessary to determine if their proposal is of clinical utility in blood. In contrast to the work that we present, the lipidomic profile was used, using a sample in peripheral blood, which could be useful for accessibility. It is relevant that our experimentation includes a validation or blind test with 30% of the data,

which the trained model is totally unaware of. This allows reinforcing the performance of the model and visualizing the behavior it will have in real life, simulating that it is subjected to a different and unknown population.

5. Conclusions

The information presented regarding the lipids capable of discriminating between babies who died or not due to SIDS provides valuable information to the state of the art. To our knowledge, it is the first work that reports the lipidomic profile in babies who die from SIDS.

The 21 characteristics used for the predictive models are potential biomarkers that are capable of predicting together (multivariate model) if a patient is likely to die from this cause. It is encouraging because lines of research emerge aimed at obtaining answers about what really causes death, so that in the future, an accurate risk profile can be obtained to prevent more deaths.

There are a few studies that used metabolomic data from children who died of SIDS to predict susceptibility to SIDS death [12,13]. It is invasive to obtain tissue samples, but in contrast, this study has the potential to use lipids that were obtained in blood. The values obtained in the blind test with 21 characteristics were very good, achieving a maximum AUC of 0.9 with the RF model, which could be used as a prediction method, and a computer tool for detect potential cases, allowing the capture and follow-up of patients with this condition.

However, it presents some limitations, such as the imbalance of the data and the small size of the sample, which will be possibly favored by a larger cohort.

An exhortation is also required from the health institutions to carry out a follow-up in this age group in order to make databases that allow obtaining relevant information, preferably in a larger cohort. A second approach that would be interesting is to analyze the 21 selected characteristics in order to reduce them according to the definitions of biomarkers in the [56] literature. Some can be considered as potential biomarkers and be submitted to the evaluation phase, both analytical and clinical, in order to verify their relevance in living patients. In addition, the metabolic or lipidomic pathways can tentatively manage to discriminate between cases and controls, which could help clearly identify patients with a higher risk of dying from this cause.

Author Contributions: Conceptualization, K.E.V.-B. and C.E.G.-T.; Data curation, J.M.C.-P.; Formal analysis, K.E.V.-B. and C.E.G.-T.; Funding acquisition, C.E.G.-T., J.I.G.-T., J.M.C.-P. and R.S.-R.; Investigation, K.E.V.-B., C.E.G.-T. and J.M.C.-P.; Methodology, K.E.V.-B., C.E.G.-T. and J.I.G.-T.; Project administration, C.E.G.-T., J.I.G.-T., H.G.-R. and R.S.-R.; Resources, J.I.G.-T. and H.G.-R.; Software, K.E.V.-B. and M.A.S.-M.; Supervision, C.E.G.-T., J.I.G.-T., H.G.-R., J.M.C.-P. and M.A.S.-M.; Validation, K.E.V.-B. and M.A.S.-M.; Visualization, H.G.-R. and R.S.-R.; Writing—original draft, K.E.V.-B., C.E.G.-T., H.G.-R., J.M.C.-P. and M.A.S.-M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data used to support the findings of this study are public. These were download from and are available on the NIH Common Fund National Metabolomics Data Repository (NMDR) website, Metabolomics Workbench, https://www.metabolomicsworkbench.org, (accessed on 22 February 2021), where it has been assigned project ID PR000475. The data can be accessed directly through its Project DOI:10.21228/M8NS47. The acquisition of the data was supported by the NIH grant, U2C-DK119886, but is not related to this manuscript, since they were taken directly from the aforementioned repository.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Horne, R.S. Sudden infant death syndrome: Current perspectives. *Intern. Med. J.* **2019**, *49*, 433–438. [CrossRef] [PubMed]
2. Bajanowski, T.; Vege, Å.; Byard, R.W.; Krous, H.F.; Arnestad, M.; Bachs, L.; Banner, J.; Blair, P.S.; Borthne, A.; Dettmeyer, R.; et al. Sudden infant death syndrome (SIDS)—Standardised investigations and classification: Recommendations. *Forensic Sci. Int.* **2007**, *165*, 129–143. [CrossRef]
3. Baruteau, A.E.; Tester, D.J.; Kapplinger, J.D.; Ackerman, M.J.; Behr, E.R. Sudden infant death syndrome and inherited cardiac conditions. *Nat. Rev. Cardiol.* **2017**, *14*, 715–726. [CrossRef] [PubMed]
4. Tester, D.J.; Wong, L.C.; Chanana, P.; Jaye, A.; Evans, J.M.; FitzPatrick, D.R.; Evans, M.J.; Fleming, P.; Jeffrey, I.; Cohen, M.C.; et al. Cardiac genetic predisposition in sudden infant death syndrome. *J. Am. Coll. Cardiol.* **2018**, *71*, 1217–1227. [CrossRef] [PubMed]
5. Izquierdo, I.; Zorio, E.; Molina, P.; Marín, P. Principales hipótesis y teorías patogénicas del síndrome de la muerte súbita del lactante. In *Libro Blanco de la Muerte Súbita Infantil*; Asociación Española de Pediatría: Barcelona, Spain, 2013; pp. 47–60.
6. Giambelluca, S.; Verlato, G.; Simonato, M.; Vedovelli, L.; Bonadies, L.; Najdekr, L.; Dunn, W.B.; Carnielli, V.P.; Cogo, P. Chorioamnionitis alters lung surfactant lipidome in newborns with respiratory distress syndrome. *Pediatr. Res.* **2021**, *90*, 1039–1043. [CrossRef]
7. Alpay Savasan, Z.; Yilmaz, A.; Ugur, Z.; Aydas, B.; Bahado-Singh, R.O.; Graham, S.F. Metabolomic profiling of cerebral palsy brain tissue reveals novel central biomarkers and biochemical pathways associated with the disease: A pilot study. *Metabolites* **2019**, *9*, 27. [CrossRef]
8. Segers, K.; Declerck, S.; Mangelings, D.; Heyden, Y.V.; Eeckhaut, A.V. Analytical techniques for metabolomic studies: A review. *Bioanalysis* **2019**, *11*, 2297–2318. [CrossRef]
9. Holčapek, M.; Gerhard, L.; Ekroos, K. Lipidomic analysis. *Anal. Chem.* **2018**, *90*, 4249–4257. [CrossRef]
10. Ochoa, B. La lipidómica, una nueva herramienta al servicio de la salud. *Gaceta Méd. Bilbao* **2006**, *103*, 101–102. [CrossRef]
11. Villa, C.; Yoon, J.H. Multi-Omics for the Understanding of Brain Diseases. *Life* **2021**, *11*, 1202. [CrossRef]
12. Graham, S.; Chevallier, O.; Kumar, P.; Türkoğlu, O.; Bahado-Singh, R. Metabolomic profiling of brain from infants who died from Sudden Infant Death Syndrome reveals novel predictive biomarkers. *J. Perinatol.* **2017**, *37*, 91–97. [CrossRef] [PubMed]
13. Graham, S.F.; Turkoglu, O.; Kumar, P.; Yilmaz, A.; Bjorndahl, T.C.; Han, B.; Mandal, R.; Wishart, D.S.; Bahado-Singh, R.O. Targeted metabolic profiling of post-mortem brain from infants who died from sudden infant death syndrome. *J. Proteome Res.* **2017**, *16*, 2587–2596. [CrossRef] [PubMed]
14. Perrone, S.; Lembo, C.; Moretti, S.; Prezioso, G.; Buonocore, G.; Toscani, G.; Marinelli, F.; Nonnis-Marzano, F.; Esposito, S. Sudden Infant Death Syndrome: Beyond Risk Factors. *Life* **2021**, *11*, 184. [CrossRef] [PubMed]
15. Tabl, A.A.; Alkhateeb, A.; ElMaraghy, W.; Rueda, L.; Ngom, A. A machine learning approach for identifying gene biomarkers guiding the treatment of breast cancer. *Front. Genet.* **2019**, *10*, 256. [CrossRef] [PubMed]
16. Wang, J.; Yan, D.; Zhao, A.; Hou, X.; Zheng, X.; Chen, P.; Bao, Y.; Jia, W.; Hu, C.; Zhang, Z.L. Discovery of potential biomarkers for osteoporosis using LC-MS/MS metabolomic methods. *Osteoporos. Int.* **2019**, *30*, 1491–1499. [CrossRef]
17. Yilmaz, A.; Ustun, I.; Ugur, Z.; Akyol, S.; Hu, W.T.; Fiandaca, M.S.; Mapstone, M.; Federoff, H.; Maddens, M.; Graham, S.F. A Community-Based Study Identifying Metabolic Biomarkers of Mild Cognitive Impairment and Alzheimer's Disease Using Artificial Intelligence and Machine Learning. *J. Alzheimer's Dis.* **2020**, *78*, 1381–1392. [CrossRef]
18. Zheng, L.; Lin, F.; Zhu, C.; Liu, G.; Wu, X.; Wu, Z.; Zheng, J.; Xia, H.; Cai, Y.; Liang, H. Machine Learning Algorithms Identify Pathogen-Specific Biomarkers of Clinical and Metabolomic Characteristics in Septic Patients with Bacterial Infections. *BioMed Res. Int.* **2020**, *2020*, 6950576. [CrossRef]
19. Bhavsar, K.A.; Singla, J.; Al-Otaibi, Y.D.; Song, O.Y.; Zikria, Y.B.; Bashir, A.K. Medical diagnosis using machine learning: A statistical review. *Comput. Mater. Contin.* **2021**, *67*, 107–125. [CrossRef]
20. Zoabi, Y.; Deri-Rozov, S.; Shomron, N. Machine learning-based prediction of COVID-19 diagnosis based on symptoms. *NPJ Digit. Med.* **2021**, *4*, 3. [CrossRef]
21. Yadav, S.S.; Jadhav, S.M. Detection of common risk factors for diagnosis of cardiac arrhythmia using machine learning algorithm. *Expert Syst. Appl.* **2021**, *163*, 113807. [CrossRef]
22. Iqbal, M.J.; Javed, Z.; Sadia, H.; Qureshi, I.A.; Irshad, A.; Ahmed, R.; Malik, K.; Raza, S.; Abbas, A.; Pezzani, R.; et al. Clinical applications of artificial intelligence and machine learning in cancer diagnosis: Looking into the future. *Cancer Cell Int.* **2021**, *21*, 1–11. [CrossRef] [PubMed]
23. Blackburn, J.; Chapur, V.F.; Stephens, J.A.; Zhao, J.; Shepler, A.; Pierson, C.R.; Otero, J.J. Revisiting the neuropathology of sudden infant death syndrome (SIDS). *Front. Neurol.* **2020**, *11*, 594550. [CrossRef] [PubMed]
24. Galván-Tejada, C.E.; Villagrana-Bañuelos, K.E.; Zanella-Calzada, L.A.; Moreno-Báez, A.; Luna-García, H.; Celaya-Padilla, J.M.; Galván-Tejada, J.I.; Gamboa-Rosales, H. Univariate Analysis of Short-Chain Fatty Acids Related to Sudden Infant Death Syndrome. *Diagnostics* **2020**, *10*, 896. [CrossRef] [PubMed]
25. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2020.

26. NIH Common Fund's National Metabolomics Data Repository (NMDR) Website, t.M.W. Lipidomics in (SIDS) Sudden Infant Death Syndrome, Project ID PR000475. 2017. Available online: https://www.metabolomicsworkbench.org/data/DRCCMetadata.php?Mode=Project&ProjectID=PR000475 (accessed on 22 February 2021).
27. Curtis, A.E.; Smith, T.A.; Ziganshin, B.A.; Elefteriades, J.A. The mystery of the Z-score. *Aorta* **2016**, *4*, 124–130. [CrossRef]
28. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
29. Lantz, B. *Machine Learning with R: Expert Techniques for Predictive Modeling*; Packt Publishing Ltd.: Birmingham, UK, 2019.
30. RColorBrewer, S.; Liaw, M.A. *Package 'Randomforest'*; University of California, Berkeley: Berkeley, CA, USA, 2018.
31. Cox, D.R. The regression analysis of binary sequences. *J. R. Stat. Soc. Ser. (Methodol.)* **1958**, *20*, 215–232. [CrossRef]
32. Sperandei, S. Understanding logistic regression analysis. *Biochem. Med.* **2014**, *24*, 12–18. [CrossRef]
33. R Core Team. Package "Stats". The R Stats Package 2018. Available online: https://stat.ethz.ch/R-manual/R-devel/library/stats/html/00Index.html (accessed on 22 February 2021).
34. Noble, W.S. What is a support vector machine? *Nat. Biotechnol.* **2006**, *24*, 1565–1567. [CrossRef]
35. Patle, A.; Chouhan, D.S. SVM kernel functions for classification. In Proceedings of the 2013 International Conference on Advances in Technology and Engineering (ICATE), Mumbai, India, 23–25 January 2013; pp. 1–9.
36. Meyer, D.; Dimitriadou, E.; Hornik, K.; Weingessel, A.; Leisch, F.; Chang, C.C.; Lin, C. Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), T.W. [R package e1071 version 1.6-7]. Comprehensive R Archive Network (CRAN), 2014. Available online: http://www2.uaem.mx/r-mirror/web/packages/e1071/ (accessed on 22 February 2021).
37. Bayes, T.L., III. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFR S. *Philos. Trans. R. Soc. Lond.* **1763**, *53*, 370–418.
38. Mann, H.B.; Whitney, D.R. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.* **1947**, *18*, 50–60. [CrossRef]
39. MacFarland, T.W.; Yates, J.M. Chapter 4. Mann–Whitney U Test. In *Introduction to Nonparametric Statistics for the Biological Sciences Using R*; Springer International Publishing: Cham, Switzerland, 2016. [CrossRef]
40. R Core Team. Wilcoxon Rank Sum and Signed Rank Tests. 2011. Available online: https://stat.ethz.ch/R-manual/R-devel/library/stats/html/wilcox.test.html (accessed on 22 February 2021).
41. Refaeilzadeh, P.; Tang, L.; Liu, H. Cross-Validation. In *Encyclopedia of Database Systems*; Liu, L., Özsu, M.T., Eds.; Springer: Boston, MA, USA, 2009. [CrossRef]
42. Kuhn, M. Building predictive models in R using the caret package. *J. Stat. Softw.* **2008**, *28*, 1–26. [CrossRef]
43. Hoo, Z.H.; Candlish, J.; Teare, D. What is an ROC curve? *Emerg. Med. J.* **2017**, *34*, 357–359. [CrossRef] [PubMed]
44. Domínguez, E.; González, R. Análisis de las curvas receiver-operating characteristic: Un método útil para evaluar procederes diagnósticos. *Rev. Cuba. Endocrinol.* **2002**, *13*, 169–176.
45. Zhu, W.; Zeng, N.; Wang, N. Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS implementations. *NESUG Proc. Health Care Life Sci.* **2010**, *19*, 67.
46. Narkhede, S. Understanding auc-roc curve. *Towards Data Sci.* **2018**, *26*, 220–227.
47. Baratloo, A.; Hosseini, M.; Negida, A.; El Ashal, G. Part 1: Simple definition and calculation of accuracy, sensitivity and specificity. *Arch. Emerg. Med.* **2015**, *3*, 48–49.
48. Kuhn, M. Caret: Classification and regression training. *Astrophys. Source Code Libr.* **2015**, ascl1505.
49. Robin, X.; Turck, N.; Hainard, A.; Tiberti, N.; Lisacek, F.; Sanchez, J.C.; Müller, M. pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinform.* **2011**, *12*, 77. [CrossRef]
50. Yu, L.; Liu, H. Feature selection for high-dimensional data: A fast correlation-based filter solution. In Proceedings of the 20th International Conference on Machine Learning (ICML-03), Washington, DC, USA, 21–24 August 2003; pp. 856–863.
51. Bommert, A.; Sun, X.; Bischl, B.; Rahnenführer, J.; Lang, M. Benchmark for filter methods for feature selection in high-dimensional classification data. *Comput. Stat. Data Anal.* **2020**, *143*, 106839. [CrossRef]
52. Hishikawa, D.; Hashidate, T.; Shimizu, T.; Shindou, H. Diversity and function of membrane glycerophospholipids generated by the remodeling pathway in mammalian cells. *J. Lipid Res.* **2014**, *55*, 799–807. [CrossRef]
53. Farooqui, A.A.; Horrocks, L.A.; Farooqui, T. Glycerophospholipids in brain: Their metabolism, incorporation into membranes, functions, and involvement in neurological disorders. *Chem. Phys. Lipids* **2000**, *106*, 1–29. [CrossRef]
54. Castro-Gómez, P.; Garcia-Serrano, A.; Visioli, F.; Fontecha, J. Relevance of dietary glycerophospholipids and sphingolipids to human health. *Prostaglandins Leukot. Essent. Fat. Acids* **2015**, *101*, 41–51. [CrossRef] [PubMed]
55. Farooqui, A.A.; Horrocks, L.A. *Glycerophospholipids in the Brain: Phospholipases A2 in Neurological Disorders*; Springer Science & Business Media: New York, NY, USA, 2006.
56. Califf, R.M. Biomarker definitions and their applications. *Exp. Biol. Med.* **2018**, *243*, 213–221. [CrossRef] [PubMed]