# Modeling and Simulation of Cell Signaling Networks for Subsequent Analytics Processes Using Big Data and Machine Learning

Bioinformatics and Biology Insights
Volume 16: 1–13
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/11779322221091739
SAGE

Máximo Eduardo Sánchez-Gutiérrez[1] and Pedro Pablo González-Pérez[2]

[1]Colegio de Ciencia y Tecnología, Universidad Autónoma de la Ciudad de México, Ciudad de México, México. [2]Departamento de Matemáticas Aplicadas y Sistemas, Universidad Autónoma Metropolitana, Unidad Cuajimalpa, Ciudad de México, México.

**ABSTRACT:** This work explores how much the traditional approach to modeling and simulation of biological systems, specifically cell signaling networks, can be increased and improved by integrating big data, data mining, and machine learning techniques. Specifically, we first model, simulate, validate, and calibrate the behavior of the PI3K/AKT/mTOR cancer-related signaling pathway. Subsequently, once the behavior of the simulated signaling network matches the expected behavior, the capacity of the computational simulation is increased to grow data (data farming). First, we use big data techniques to extract, collect, filter, and store large volumes of data describing all the interactions among the simulated cell signaling system components over time. Afterward, we apply data mining and machine learning techniques—specifically, exploratory data analysis, feature selection techniques, and supervised neural network models—to the resulting biological dataset to obtain new inferences and knowledge about this biological system. The results showed how the traditional approach to the simulation of biological systems could be enhanced and improved by incorporating big data, data mining, and machine learning techniques, which significantly contributed to increasing the predictive power of the simulation.

**KEYWORDS:** Big data techniques, modeling and simulation approach, cell signaling networks, machine learning techniques, predictive analytics

**RECEIVED:** December 14, 2021. **ACCEPTED:** March 16, 2022.

**TYPE:** Original Research

**FUNDING:** The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was partially funded by the authors' universities and the *Consejo Nacional de Ciencia y Tecnología (CONACyT)* from México. Its contents are the responsibility of the authors and do not reflect the views of the research granting bodies. The authors were responsible for the data analysis after the extraction and linkage.

**DECLARATION OF CONFLICTING INTERESTS:** The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**CORRESPONDING AUTHOR:** Pedro Pablo González-Pérez, Departamento de Matemáticas Aplicadas y Sistemas, Universidad Autónoma Metropolitana, Unidad Cuajimalpa, Avenida Vasco de Quiroga 4871, Col. Santa Fe Cuajimalpa, C.P. 05348, Ciudad de México, México. Email: pgonzalez@cua.uam.mx

## Introduction

Big data techniques are commonly applied whenever raw data is too large to be processed by a computer system. Big data also refers to when the management systems or database servers cannot provide the required data in a reasonable time due to problems with loading, searching, selecting, and saving.[1-3]

It is difficult to find a widely accepted definition of the term big data because, in the vast majority of cases, the proposed definitions for this term are dependent on the domain in which it has been used.[1] However, when we think of big data, we are referring to (1) data whose volume and complexity require more sophisticated methods for storage, retrieval, interaction, analysis, and inferences; (2) software systems whose functionality is already unsuitable for dealing with the volume and complexity of the data they must process; (3) large volumes of data that involve both structured and unstructured data, which makes their treatment much more complex; and (4) the application of powerful computational processing to highly massive and complex datasets.

A commonly accepted big data approach is the 3 "Vs" that characterize it: Volume, Velocity, and Variety.[4]

- **Volume**. It refers to the large dimensions of data generated, collected, and in many cases, constantly analyzed. Volume is precisely the characteristic that we most associate with big data; it is impossible not to think of volume when it comes to big data. Depending on its size, the volume of big data can be measured in megabytes, gigabytes, terabytes, and even petabytes.
- **Velocity**. It means the speed with which data is generated, collected, and processed. Let us think about the remarkable speed with which data is generated in applications such as search engines, the stock market, e-commerce platforms, and social networks, to name a few examples. In real-time computer systems, response time becomes an essential variable.
- **Variety**. Indicates the non-homogeneity or diversity of the data because it comes from very different origins or sources, implying that the data are of very different types, such as numeric, Boolean, categorical, nominal, ordinal, structured text, unstructured text, images, and videos among others.

However, in recent years, other "Vs" have been proposed to contribute to the definition of big data, such as Veracity and Value. The term Veracity refers to the low quality that sometimes characterizes large-scale data. In other words, the need to contend with the uncertainty in the data, mainly due to the wide variety that characterizes them derived from the different sources that generate them. The veracity problem in large-scale data commonly occurs in unstructured text data, commonly generated in sources such as social networks, emails, and chats, due to the freedom that characterizes its creation. On the contrary, the term Value refers to the importance and significance that big data can provide in making decisions that lead to companies, businesses, and institutions being much more profitable and successful.

It is undeniable that the computational simulation approach could be enhanced and improved by integrating big data techniques,[1] which would be a valuable support for the acquisition, processing, and analytics of the large volume of data that computer simulations produce continuously. Specifically, big data techniques provide a means to obtain and evaluate large-scale data produced for computational simulations, as well as to extract causal and temporal relationships between input and output patterns, which will allow us to carry out further predictions and inferences about the behavior of the simulated system.[5,6]

Biology has been one of the disciplines strongly favored with the support of computational simulation. In the last 2 decades, a wide range of computational simulations of biological systems and processes, such as protein folding, artificial foldamer design, molecular docking, and cell signaling networks, among others, have been developed based on a wide range of mathematical and computational models (see a survey of these models[7]). In particular, the modeling and simulation of cell signaling networks have ranged from continuous and discrete mathematical models[8] such as systems of differential equations or numerical methods, respectively; to computational models,[9-12] such as cellular automata,[13,14] Petri nets,[15,16] Boolean models,[17,18] rule-based systems,[19,20] multiagent systems,[12,21,22] and artificial neural networks (ANNs).[23,24]

During the execution of the simulation of the biological system, large volumes of data are generated, eg, proteomics, genomics, interactomics, and metabolomics, among others, which, once acquired, structured, and stored, will undoubtedly constitute a valuable input for the predictive analytics application, commonly based on data mining techniques.[25]

Therefore, in this piece of work, we are exploring how much the traditional approach to modeling and simulation of biological systems, specifically cell signaling networks, can be increased and improved by integrating big data and machine learning techniques.

Essentially, in this work we (1) model the PI3K/AKT/mTOR signaling network, identifying cellular compartments, signaling elements, the types of interactions between them, and the kinetic parameters and initial parameters concentrations that characterize the interactions, and the signaling elements, respectively; (2) simulate, verify, and calibrate the expected behavior of the PI3K/AKT/mTOR signaling network on the Big-Data Cellulat bioinformatics platform; (3) generate large volumes of data describing the behavior of the simulated biological system over time; and (4) apply exploratory analysis, feature selection, and analytics processes to the resulting biological dataset, to obtain new inferences and knowledge about this biological system.

We state that when the computational simulation of a biological system has been finely tuned and verified then, beyond the observed simulated behavior and the subsequent in silico experiments carried out, one of its strengths lies in the large volume of reliable biological data that it can produce. As a result, these data, through exploratory analysis and analytics process, can produce new inferences and knowledge about the simulated biological system.

## Material and Methods

### *The PI3K/AKT/ mTOR signaling pathway*

Intracellular signaling is one of the essential molecular mechanisms for controlling cell activity, and it is involved in almost all cell functions, including cell division, growth, differentiation, and death. Cancer progression, malignancy, and treatment resistance are all influenced by signaling pathways. The intracellular signaling pathways linked with NF-kB (nuclear factor kappa B), TGF-β (transforming growth factor beta), Notch, and PI3K/AKT/mTOR are the most commonly altered in cancer.

The PI3K/AKT/mTOR signaling pathway is engaged in many tasks relevant to cancer biology, including cellular proliferation, survival, migration, angiogenesis, and apoptosis,[26-28] making it one of the most critical processes in cancer growth. The research of anticancer targets working through the PI3K/AKT/mTOR signaling pathway requires a comprehensive understanding of this signaling pathway: the characteristics of its signaling elements, the complex interactions that occur between them—signal amplification, activation, deactivation, phosphorylation, dephosphorylation, complex formation, and several others—and the global behaviors that ensue; which requires the use of approaches such as systems biology, big data, and data mining. It is critical for cancer treatment to understand the activity of PI3K/AKT/mTOR and how it interacts with other pathways that are regulated by the presence of specific molecules.

In several types of cancers, including brain, breast, ovarian, and renal carcinomas, unregulated activation of the PI3K/AKT/mTOR signaling pathway contributes to cellular change and tumor growth. The PI3K/AKT/mTOR pathway is important for a variety of cellular processes[29] because it contains a complicated signaling mechanism including 3 key modulators proteins: PI3K (Phosphoinositide-3-kinase), AKT (Serine/threonine kinase, also known as protein kinase B), and mTOR (Serine/threonine kinase, mammalian Target of Rapamycin).

### *Big-Data Cellulat: the cell signaling network simulator*

The Big-Data Cellulat platform[12,30,31] was conceived as a computer simulator for cellular signal transduction systems.

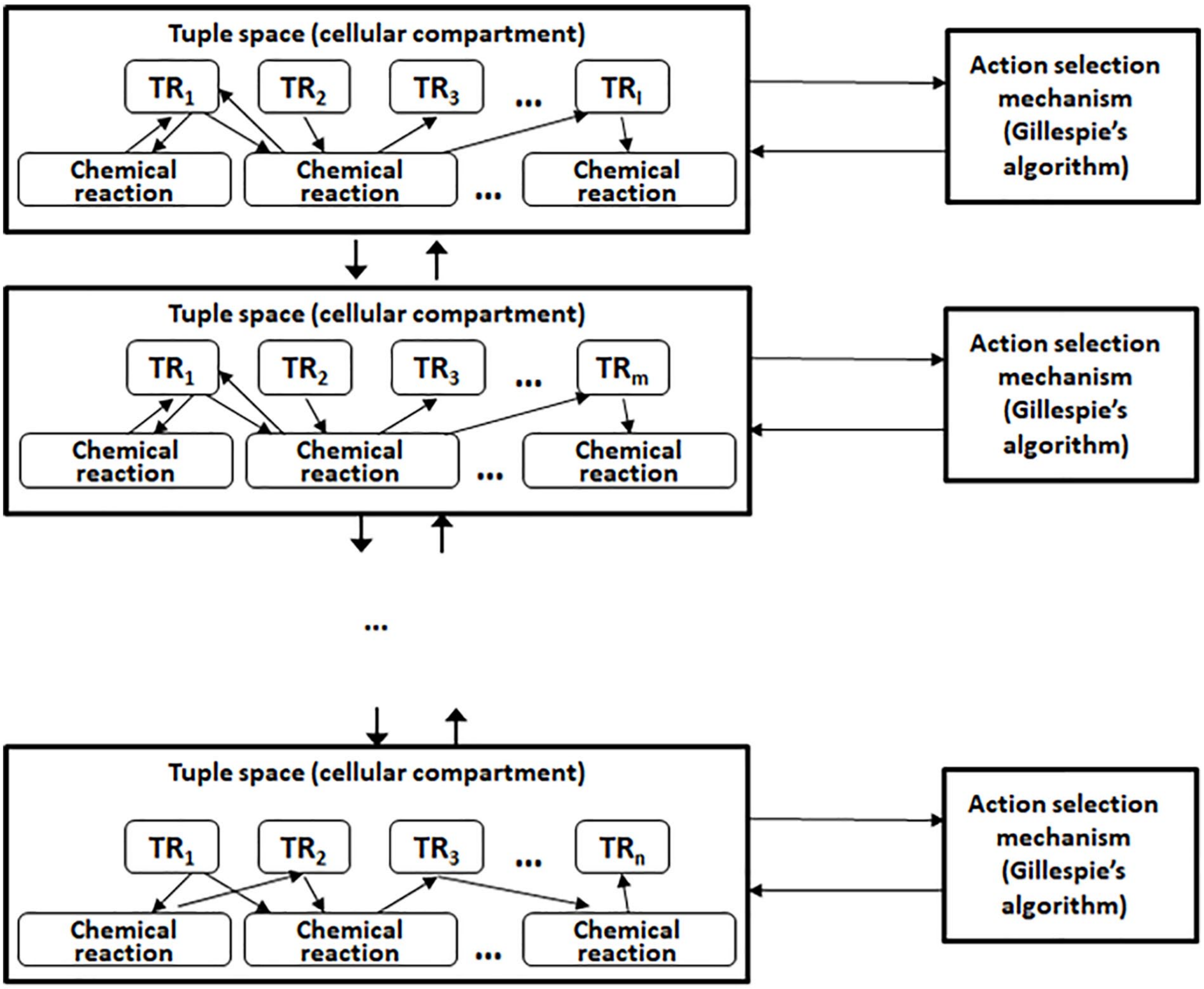
Tuple space (cellular compartment)
TR$_1$
TR$_2$
TR$_3$
...
TR$_l$
Chemical reaction
Chemical reaction
...
Chemical reaction
Action selection mechanism (Gillespie's algorithm)
Tuple space (cellular compartment)
TR$_1$
TR$_2$
TR$_3$
...
TR$_m$
Chemical reaction
Chemical reaction
...
Chemical reaction
Action selection mechanism (Gillespie's algorithm)
...
Tuple space (cellular compartment)
TR$_1$
TR$_2$
TR$_3$
...
TR$_n$
Chemical reaction
Chemical reaction
...
Chemical reaction
Action selection mechanism (Gillespie's algorithm)

**Figure 1.** Use of tuple spaces to represent cell compartments, reactants, and chemical reactions involved in cell signaling. Note that the selection and execution of chemical reactions are coordinated by an action selection mechanism based on the Gillespie algorithm.

The simulator is based on a model that integrates (1) the concept of tuple space[12,32,33] for the representation and interaction of chemical reactions and reactants and (2) an action selection mechanism based on the Gillespie algorithm,[34,35] for the selection and execution of chemical reactions. The joint use of these 2 approaches allows Big-Data Cellulat to exhibit a series of key characteristics required to simulate cell signaling systems and the consequent in silico experimentation. On one hand, the representation based on tuple spaces provides the simulation with characteristics such as multicompartmentalization, localization, and topology; on the other hand, the selection and execution of chemical reactions based on the Gillespie algorithm provide the simulation with synchronization, timing, and a selection based both on the rate/affinity of the chemical reaction and on the availability of the reactants.

*Representation of the chemical reactions and reactants.* A tuple is an ordered collection of information or knowledge, and as a knowledge representation, tuples aid in representing the chemical reactions and reactants. In a tuple space, the interaction and synchronization between functions, procedures, objects, programs, and even intelligent agents, occur through reading, modifying, writing, and destroying tuples in the shared tuple space.[32,33] Based on these considerations, the translation of the structures and elements involved in cell signaling to abstractions of the tuple spaces is shown in Figure 1 and described in Table 1.

*Selection and execution of the chemical reactions.* As previously mentioned, in Big-Data Cellulat, the selection and execution of the chemical reactions are carried out by an action selection mechanism based on the Gillespie algorithm,[34,35] which selects the next reaction to occur considering a random number and the propensity function of the reaction. The propensity function is calculated based on the rate/affinity of the reaction and the molar availability/molar requirement of the reactants involved in this reaction. The core of the action selection mechanism can be summarized in the following steps and expressions (1) to (3).

**Table 1.** Structures and elements involved in cell signaling translation to abstractions of the tuple spaces.

| STRUCTURES AND COMPONENTS INVOLVED IN CELL SIGNALING | TUPLE SPACE MODEL ABSTRACTIONS |
|---|---|
| Tissue, cells, extracellular milieu, and intracellular compartments (ie, cell membrane, cytosol, nucleus, mitochondrion, among others) | Tuple space |
| Signaling components such as membrane receptors, proteins, enzymes, transcription factors, and genes | Sets of chemical reactions (which can be seen as simple agents) |
| Signaling molecules (ie, ligands, second messengers, substrates) | Reactants and their concentration values represented as tuples in the tuple space |

1. Calculation of the rate for each chemical reaction *j*.

$$Rate_j = RateConstant^* \prod_{i=1}^{k} \left( \frac{Mol_i}{reqMol_i} \right) \tag{1}$$

where *RateConstant* is the reaction rate constant, $Mol_i$ is the number of available molecules of reactant $i$, $1 \leqslant i \leqslant k$, and *reqMoli* is the number of molecules required of reactant $i$, $1 \leqslant i \leqslant k$.

2. Selection of the next chemical reaction to run.

$$\psi \leq \frac{\sum_{i=1}^{n} Rate_i}{RTot} \tag{2}$$

where $\psi$ is a random number, $0 \leqslant \psi \leqslant 1$ and *RTot* is the summation of the rates ($Rate_i$) of all reactions.

3. Determination of the delay (suspension) between the last reaction executed and the next reaction to be executed.

$$Stop_{time} = \frac{-\ln(\tau)}{RTot} \tag{3}$$

where $\tau$ is a random number, $0 \leqslant \tau \leqslant 1$.

*The main characteristics and functionality of Big-Data Cellulat simulator.* As a computational simulator of cell signaling systems, Big-Data Cellulat exhibits characteristics that are crucial when trying to emulate the structure and behavior of this type of complex biological systems, such as compartmentalization, localization, topology, interaction, coordination, synchronization, timing, and selection and execution of chemical reactions considering their rate/affinity. As previously mentioned, these characteristics emerge from the joint use of (1) a tuple space model for the representation and interaction of the structures, elements, and components involved in cell signaling and (2) coordination and action selection mechanism based on the Gillespie algorithm, for the selection and execution of chemical reactions. At this point, it is necessary to note that the functionality exhibited by the Big-Data Cellulat simulator can be described in terms of the characteristics mentioned above. That is, each of these characteristics constitutes in itself a feature that the simulation tool provides to the user during the phases: (1) creation of the simulation; (2) execution, calibration, and validation/verification of the simulation; (3) execution of in silico experiments; and (4) production and recording of massive data for intelligent data analysis tasks.

### *Farming big cell signaling data*

As pointed out by Tolk,[1] among the big-data methods closely related to the modeling and simulation approach are data farming and crowdsourcing. Both methods are beneficial when applied to steps of traditional modeling and simulation studies. In particular, data farming uses computational simulations (in silico experiments) to grow data. Once the data are produced and stored—in this case by the Big-Data Cellulat simulator—it can be analyzed using various techniques and models, such as data mining and machine learning, to discover causal relationships between them. When the Big-Data Cellulat simulator is used, data farming takes place once the simulation is launched, encompassing the following actions:

1. Selection/filtering of the features (signaling elements) required for the dataset integration.
2. Selection of the sampling factor (*K* milliseconds, seconds or minutes, with *K* integer, *K* > 0) for data recording.
3. Identification of the path, file name, and extension where the data will be stored.

### *Exploratory data analysis and feature selection techniques*

The purpose of the exploratory data analysis is to generate as many insights and information about the data as possible and find any problems in the dataset. One of the most common issues found in datasets is missing values. Two frequently used techniques to handle missing values in a dataset are dropping rows or columns and replacing missing values with central tendency values such as mean, median, and mode. Deleting rows or columns with missing values may produce a model that works poorly if the percentage of missing values is excessive compared to the complete dataset. On the contrary, inputting missing values prevent data loss but do not factor the covariance between features.

Another common issue concerning datasets is the unequal distribution of classes within a dataset, known as data imbalance. Some techniques can be used to solve the class imbalance

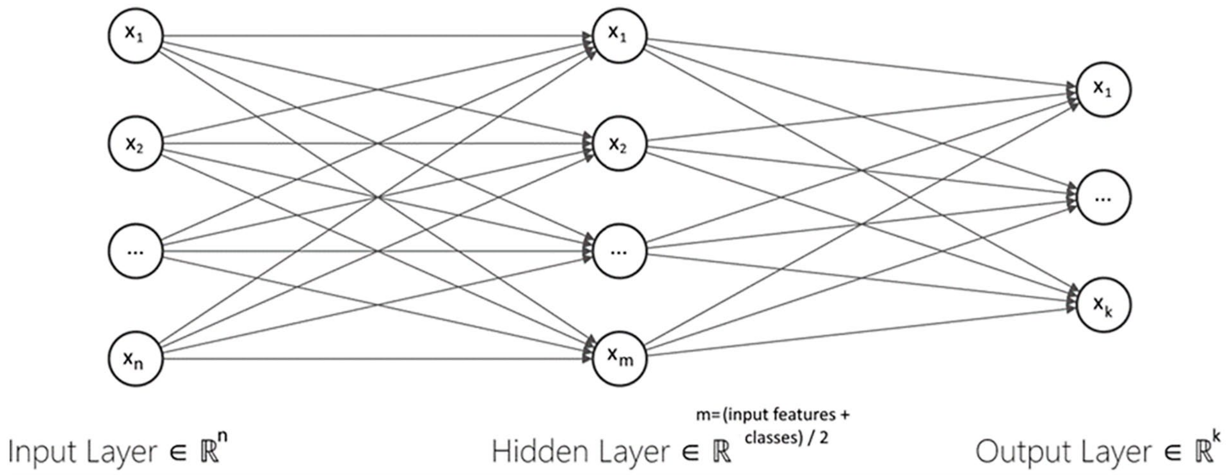

**Figure 2.** Multilayer perceptron model. The input layer contains 77 neurons corresponding to the dimension of the input vectors (ie, the total number of signaling elements in the PI3K/AKT/mTOR pathway). The hidden layer contains (# input features + # classes)/2 neurons. If all features are considered, it would have 41 neurons. Finally, the output layer contains 6 neurons corresponding to the same number of classes representing cell states or possible combinations.

problem; resampling by oversampling or undersampling[36] and ensemble methods.[37] Oversampling can be carried out by generating as many synthetic samples as needed, selecting the most common value in the class and repeating it, or repeating a randomly selected value from the smallest class. In contrast, undersampling is a technique that decreases the number of samples of the most significant class down to the smallest class size. These 2 techniques can be combined to oversample the minority class and undersample the majority class. On the contrary, ensemble methods typically use boosting or bagging to build several estimators on a different randomly selected subset of data.

One advantage of the Big-Data Cellulat simulator is that it can farm data by computational simulations (in silico experimentation). In the context of data imbalance, artificial generation of data points is unnecessary because the simulation can be run multiple times and, consequently, join the resulting datasets. In this case, the class imbalance can be dealt with undersample techniques to reduce the majority classes. In general, raw datasets contain various data types, including numerical and categorical information. Feature engineering deals with these heterogeneous datatypes using various techniques that convert different data types to numerical vectors.[38] For example, to encode categorical or numerical features, one can use Dummy Encoding, Count Encoder, One-Hot Encoder, and idxmax, among others. Similarly, feature binning converts continuous to categorical variables.

### *Predictive process*

Once the raw dataset is preprocessed, and the class information is encoded, the data are ready to be fed to a predictive model. In this work, we choose a multilayer perceptron (MLP)[39] to predict the cellular state or states that should characterize the cell, given a particular activation/deactivation configuration of the signaling elements that make up the network. An MLP can be implemented as a classifier because it finds the most appropriate boundary between 2 or more classes. Hence, it may discern the structural differences between 2 or more given classes, identify the space that separates each one, and determine the likelihood of a given data point belonging to a particular class. An MLP is a neural network connecting multiple neurons or perceptrons, partitioned into the input layer, the hidden layer, and the output layer. The neurons compose a directed acyclic graph, meaning that the paths connect nodes in layers from one layer to the next, as shown in Figure 2. Each neuron, excluding the input ones, has a nonlinear activation function, a bias, and connecting weights which the MLP train by backpropagation in a supervised learning fashion[40] so that the error value can be updated in a much more successful way.

When developing a neural network model, 3 stages are needed before its deployment: (1) dataset preprocessing, (2) performing feature engineering, and (3) dividing the dataset into training and testing sets using a cross-validation strategy. The input dataset to the machine learning model usually requires partitioning the data into training and test sets. Data belonging to the training set contains a known output or label, from which the model learns to generalize to other data. On the contrary, the test set is used to test our model's prediction capabilities. In this work, we perform a cross-validation schema to split the dataset by partitioning the available data into 3 sets (see Figure 3).

### *Methodological approach*

The methodological approach followed in this work integrates the key aspects of traditional modeling and simulation with current big data, data mining, and machine learning techniques, involving the following activities:

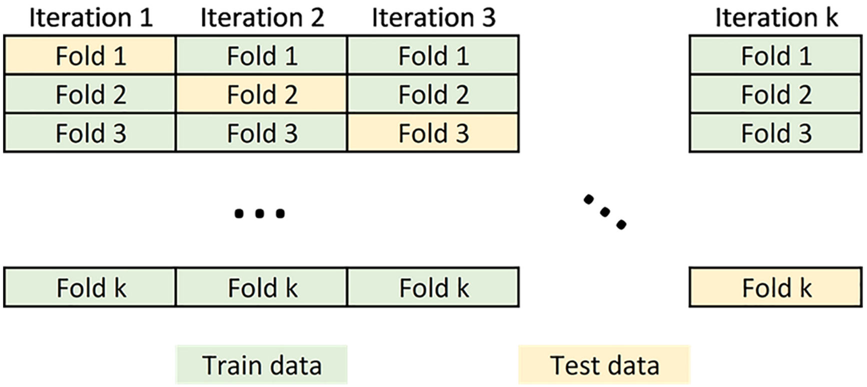

**Figure 3.** Cross-validation schema. In the *k*-fold cross-validation, the training set is split into *k* smaller sets. A model is trained using *k* – 1 of the folds as training data; the resulting model is validated on the test set to compute a performance measure. The performance measure reported by *k*-fold cross-validation is then the average of the values computed in the loop.

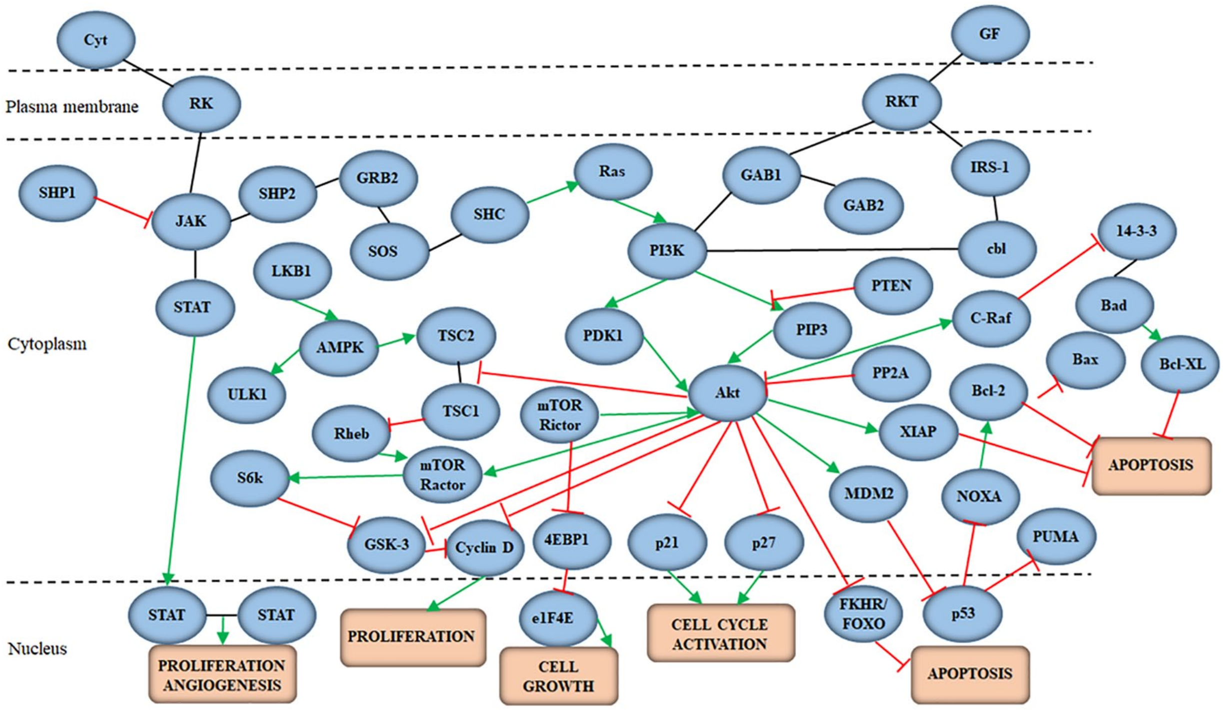

**Figure 4.** The resulting model of the PI3K/AKT/mTOR cell signaling pathway. The model involves the main interactions in this pathway, from the binding of ligands to transmembrane receptors to the triggering of cellular states, characteristic of breast cancer cells. Green arrows represent activation interactions, red lines represent inhibition interactions, and black lines represent compound formation.

1. **Modeling**. Modeling the PI3K/AKT/mTOR signaling network, identifying cellular compartments, signaling elements, the types of interactions between them, and the kinetic parameters and initial concentrations that characterize the interactions and the signaling elements, respectively.
2. **Creation of the computational model (simulation)**. Assembly in Big-Data Cellulat of cell structures (cell compartments, cells, tissues), chemical reactions with their kinetic parameters, and reactants with their initial molar concentration value.
3. **Simulation, validation, and calibration**. Simulation, verification, and validation of the expected behavior of the PI3K/AKT/mTOR signaling network in the Big-Data Cellulat bioinformatics platform.
4. **Data farming**. Generation of big data describing the behavior of the simulated biological system over time.
5. **Intelligent data analysis**. Application of data mining and machine learning techniques to the resulting biological dataset to obtain new inferences and knowledge about this biological system.

## Results and Discussion

### *The modeling of the PI3K/AKT/mTOR signaling network*

The main results obtained in the modeling phase are illustrated in Figure 4 and described in Table 2. Figure 4 shows the resulting model of the PI3K/AKT/mTOR signaling network-integrated from segments, cascades, particular types of interactions, as well as other theoretical-experimental aspects reported in the

**Table 2.** Examples of chemical reactions are defined as part of the modeling of the PI3K/AKT/mTOR signaling pathway.

| REACTION | REACTANTS | INITIAL CONC. (μMOL) | $K_M$ (μMOL) | $V_{MAX}$ (μMOL/μL/SEG) | $V_0$ |
|---|---|---|---|---|---|
| Cyt + RK → CytRK | Cyt<br>RK | 0.1<br>0.25 | 34.2 | 7.6 | $2.22 \times 10^{-5}$ |
| CytRK + JAK → CytRKJAK* | JAK<br>Cyt<br>RK | 0.012<br>0.0001<br>0.25 | 34.2 | 7.6 | $2.22 \times 10^{-5}$ |
| CytRKJAK* + STAT → STAT* | STAT<br>Cyt<br>RK | 0.4<br>0.0001<br>0.25 | 74.1 | 49 | $6.61 \times 10^{-5}$ |
| RAS* + PI3K → PI3K* | PI3K<br>RAS | 0.9<br>0.8 | 53.4 | 49 | 0.0915 |
| PIP3* + Akt → Akt* | PIP3<br>Akt | 0.27<br>0.2 | 1.1 | 22.1 | 4.3554 |
| PDK1* + Akt → Akt* | PDK1<br>Akt | 1.0<br>0.2 | 36 | 22.3 | 0.6027 |
| Akt* + p27* → p27 | Akt<br>p27 | 0.2<br>0.27 | 7.8 | 8.4 | 0.2810 |
| Akt* + FKHR* + FOXO* → FKHR/FOXO | Akt<br>FKHR<br>FOXO | 0.2<br>0.4<br>0.4 | 74.1 | 49 | 0.2630 |
| FKHR/FOXO → Apoptosis inhibition | FKHR/<br>FOXO | 0.4 | 74.1 | 49 | 0.2630 |
| STAT* → Proliferation/Angiogenesis | STAT* | 0.4 | 74.1 | 49 | 0.2630 |
| p27 → Cell cycle activation | p27 | 0.27 | 7.8 | 8.4 | 0.2810 |

specialized literature on this antiapoptotic signaling pathway and its role in cancer. In the representation of the signaling network illustrated in Figure 4, the nodes represent signaling elements such as membrane receptors, proteins, and transcription factors; the arcs establish the different types of interaction that occur between the signaling elements such as activation, inhibition, compounding, among others, while the rounded edge rectangles correspond to the final cell states achievable from specific activation/inhibition combinations of signaling elements. Note that the primary cellular compartments involved in intracellular signaling have also been identified. On the contrary, Table 2 provides examples of the reactions that formalize the interactions between the signaling elements illustrated in Figure 4. As shown in Table 2, each reaction is characterized by its kinetic parameters and the concentration micromolar initial of the reactants involved; all these parameters are required in the subsequent simulation creation phase. It should be noted that the global signaling network illustrated in Figure 4 involves more than 60 reactions.

### *The simulation of the PI3K/AKT/mTOR signaling network*

The behavior of the simulation of the PI3K/AKT/mTOR signaling pathway over time (on a millisecond scale) is captured in the snapshots illustrated in Figure 5. All in silico experiments carried out during this phase aimed to validate, verify, and calibrate the simulated biological system, which was facilitated by running the simulation in phases of incremental complexity. That is, the signaling pathway was split into the following signaling segments identified from a biological approach: (1) from the activation of transmembrane receptors to the activation of proteins and the formation of compounds in the juxtamembrane region, (2) from the activation of transmembrane receptors to the activation of key proteins and the formation of compounds in the cytoplasm, (3) from the activation of transmembrane receptors to the activation of transcription factors in the nucleus, and (4) from the activation of transmembrane receptors to the triggering of final cellular processes.

The simulation verification, validation, and calibration processes aim to reduce the error between the simulated micromolar concentration values of the target signaling elements and their concentration values of these signaling elements observed in the real biological system. The simulation validation was based on the analysis of differences between simulated concentration values and measured concentration values, using statistical indices such as the mean bias error (MBE), the mean absolute error (MAE), the mean square error (MSE), and the root mean square error (RMSE).

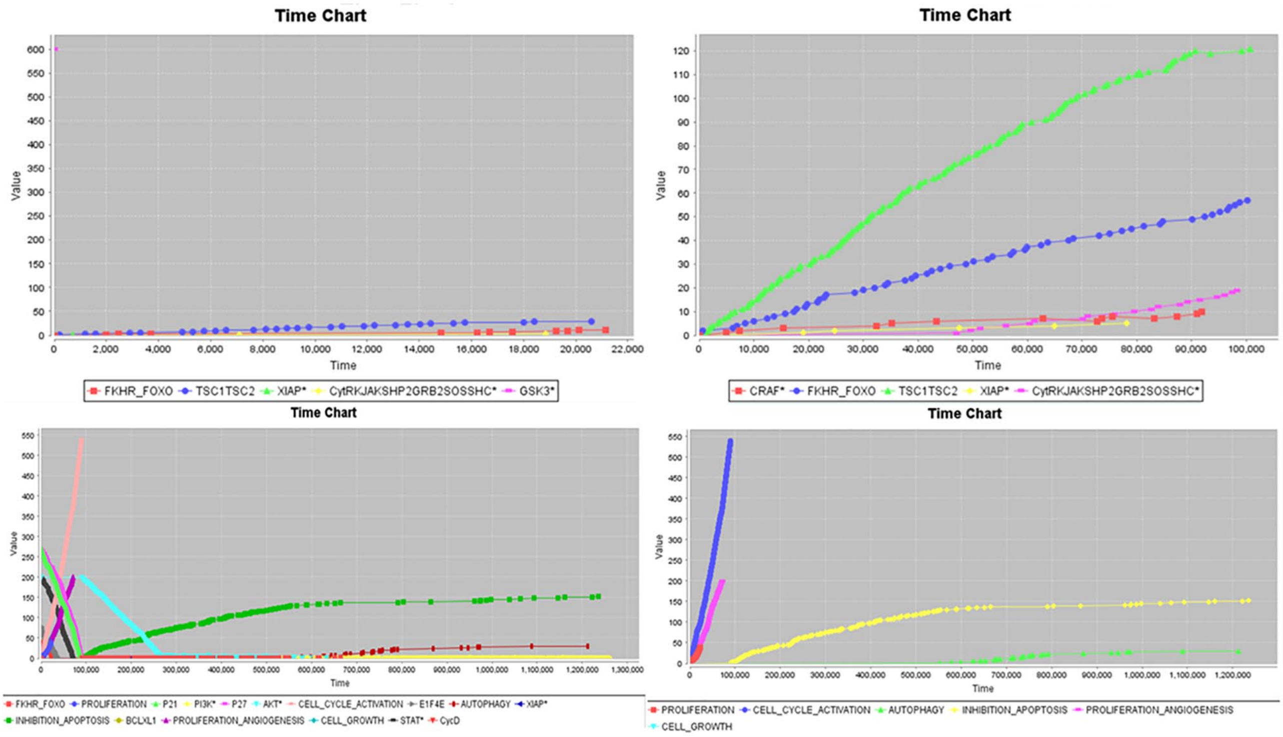

**Figure 5.** Simulation of the PI3K/AKT/mTOR signaling network in Big-Data Cellulat simulator.

**Table 3.** Characteristics of the PI3K/AKT/mTOR signaling dataset resulting from the data farming process.

| BIOLOGICAL DATASET | NUMBER OF FEATURES (SIGNALING ELEMENTS) | NUMBER OF INSTANCES | NUMBER OF CLASSES (CELLULAR PROCESSES) |
|---|---|---|---|
| PI3K/AKT signaling network | 77 | 35574 | 6 |

### *The big cell signaling dataset produced by the computer simulator*

As previously described, the main product resulting from the data farming stage is a large volume of input-output patterns produced by in silico experiments carried out by Cellulat bioinformatics framework. In this case, the product is the simulation of the behavior of the PI3K/AKT/mTOR signaling network from different concentration settings of its signaling elements.

Table 3 shows the characteristics of the resulting dataset in terms of the number of features, number of instances, and number of classes. The dataset comprises 35574 instances characterized by 77 attributes and 6 classes in which instances are classified. The 77 attributes are the input patterns produced by in silico experiments and are represented by the concentration value of the signaling elements such as receptors, key proteins, PI3K/Ras inhibitors, anti-apoptotic proteins, pro-apoptotic proteins, and tumor suppressor proteins. On the other hand, classes represent the cellular states—eg, Autophagy, Proliferation, Inhibition Apoptosis, Cell Growth, Proliferation Angiogenesis, and Cell Cycle Activation—to which the cell could be carried, depending on a particular activation/inhibition configuration exhibited by the signaling elements. Note that an input pattern could be classified in more than one class. In other words, the classes are not mutually exclusive. Table 4 shows the number of instances in each class for each sampling period.

As mentioned earlier, machine learning requires a dataset with which the learning process can be carried out. First, this dataset needs to go through a preprocessing stage, including data cleansing, to transform it into a format that a machine learning algorithm can understand. In our dataset, as shown in Table 4, the signaling pathway identifies 6 classes in which the instances are grouped as Autophagy, Proliferation, Inhibition Apoptosis, Cell Growth, Proliferation Angiogenesis, and Cell Cycle Activation. The discrepancy between instances reported in Tables 3 and 4 is due to the classes not being mutually exclusive. To prevent that one instance can be assigned to more than one class, the initial representation of the signaling database was transformed to a suitable scheme that would allow its processing by machine learning algorithms, as shown in Table 5. With this new assignment of patterns to classes, the number of instances grouped in each of them underwent the update shown in Table 6 and Figure 6.

Observe in Figure 6 that the number of instances is imbalanced; the majority class is about 20.7 times the minority class. This class imbalance can lead to models biased toward the majority class, causing the wrong classification of the minority

**Table 4.** Classes and number of instances for 3 PI3K/AKT/mTOR signaling datasets generated from 3 different sampling rates: 50, 100, and 500 milliseconds.

| CLASS | NUMBER OF INSTANCES | | | |
|---|---|---|---|---|
| | 50 MS | 100 MS | 500 MS | TOTAL |
| Autophagy | 10 565 | 4295 | 1056 | 15 916 |
| Proliferation | 19 104 | 7766 | 1910 | 28 780 |
| Inhibition apoptosis | 21 256 | 8641 | 2125 | 31 959 |
| Cell growth | 21 976 | 8933 | 2197 | 33 106 |
| Proliferation angiogenesis | 23 103 | 9391 | 2310 | 34 804 |
| Cell cycle activation | 23 614 | 9599 | 2361 | 35 574 |

**Table 5.** Adopted classification scheme.

| ATTRIBUTES VALUES | PREVIOUS CLASSES | | | | | | NEW CLASS |
|---|---|---|---|---|---|---|---|
| | AUTOPHAGY | PROLIFERATION | INHIBITION APOPTOSIS | CELL GROWTH | PROLIFERATION ANGIOGENESIS | CELL CYCLE ACTIVATION | |
| | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | 0 | 1 | 1 | 1 | 1 | 1 | 2 |
| . . . | 0 | 0 | 1 | 1 | 1 | 1 | 3 |
| | 0 | 0 | 0 | 1 | 1 | 1 | 4 |
| | 0 | 0 | 0 | 0 | 1 | 1 | 5 |
| | 0 | 0 | 0 | 0 | 0 | 1 | 6 |

**Table 6.** Characteristics of the signaling dataset resulting from the data farming process.

| BIOLOGICAL DATASET | NUMBER OF FEATURES (SIGNALING ELEMENTS) | NUMBER OF INSTANCES | NUMBER OF CLASSES (1 ATTRIBUTE) |
|---|---|---|---|
| Signaling network | 77 | 35 574 | 6 |

class. To alleviate this problem, we are presented with 2 options: oversample the minority classes or undersample the majority classes. As the dataset comes from a data farming stage of the patterns produced by in silico experiments carried out by the Cellulat bioinformatics framework, the number of samples that we can obtain is only restricted by the framework processing time; this presents us with the opportunity to use the undersampling technique. As stated in Table 6, the dataset is composed of 77 descriptors, and one attribute representing the class, the whole relation of features (signaling elements) and interactions are shown in Figure 4.

### *Results of the exploratory data analysis and feature selection*

As previously stated, because the number of instances is imbalanced, in this work, we handle the class imbalance by undersampling the majority classes avoiding the need to generate samples from the same dataset artificially. This technique produces a random subsample of a dataset by removing random observations of the majority classes, and the redistribution of samples is shown in Figure 7.

Once the dataset was balanced, we explored a set of feature selection techniques to rank the features according to their saliency to get an idea of the importance of the features. These techniques consider data variance, chi-square stats, feature ranking with recursive cross-validated feature elimination, a linear model with iterative fitting, a meta estimator that fits randomized decision trees, and Pearson correlation. An excerpt of the 77 ranked features is available in Table 7. As each technique ranks the features differently, Table 7 shows the relative importance of the features for each technique. It is important to note that if we group the features in the top 10, 20, . . ., some features are found to belong in the same tier across selection techniques, eg, PI3K*, AKT, BAX*, AMPK*, among others.

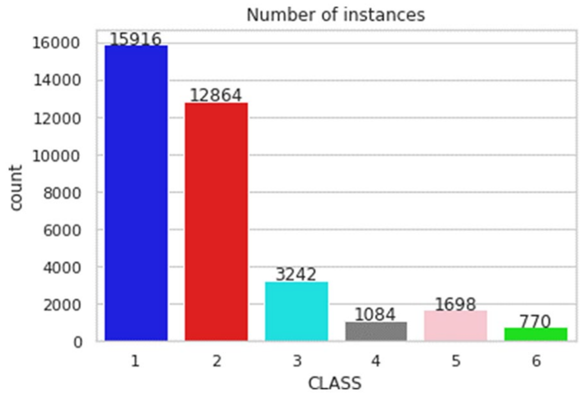

**Figure 6.** Signaling dataset produced by the simulator. There is an unequal class distribution in the dataset; the minority class is about 20 times smaller than the majority class, leading to a poor model's prediction. The reassignment of instances to the new proposed classes is depicted in Table 5.

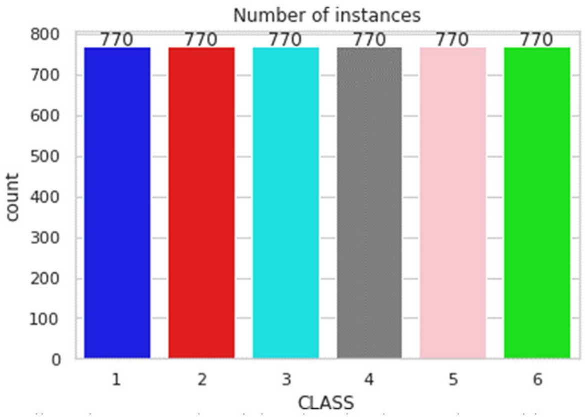

**Figure 7.** Signaling dataset produced by the simulator. The problem of an unequal class distribution in the training dataset was solved by undersampling the majority classes.

After ranking the features, we can select the ones that meet specific criteria, eg, the ones that explain 80% of the total variance, and in general, the features that are the most likely to be class-dependent and therefore more relevant for classification. Below we present the prediction results produced by the MLP model after applying several selection techniques using (1) the entire dataset (77 features/descriptors), (2) the domain expert-suggested feature selection (12 features/descriptors), (3) the complete set of primary descriptors identified by each technique, and finally, (4) the complete set of primary descriptors identified by each technique enriched with a subset of features according to domain experts.

### *Results of the predictive process on the big cell signaling dataset*

The results of the 4 experiments that explored the effects of the different selection techniques are shown in Table 8. To have a baseline to compare against, we use the entire dataset without any feature selection, ie, the 77 features. This first experiment achieved an accuracy of 96.15%. In contrast, the second experiment used the 12 features suggested by the domain expert, obtaining only 93.64% accuracy. The 12 suggested features (signaling elements) in this case were CycD, BCL2, E1F4E, P27, BCLXL1, STAT*, XIAP, PI3K, AKT, P21, RAS, and FKHR-FOXO (see the whole PI3K/AKT/mTOR signaling pathway in Figure 4). Regarding the number of features, only 15.6% of the features were used in the second experiment; this means that there is room for improvement in the number of features selected and accuracy. This also means that the features suggested by the domain expert may not be the best ones to describe the data.

We ran a third group of experiments to investigate the effect of selecting different subsets of features by different techniques. The results of such a group of experiments are listed in the *Selected features* column of Table 8; in this column, we can see that the *Recursive feature elimination* technique yields a better accuracy rate using only 23 of the 77 original features.

Finally, the results of the *Selected* and *domain knowledge* features experiment in Table 8 show that accuracy wise, not only it is not beneficial to add the suggested domain features, but if we compare the cardinality (number of features) of both experiments, we can also see an increment. In this regard, note that the Pearson correlation technique worsens the accuracy by adding only one more expert-suggested feature set. Contrarily, the accuracy of the low variance technique significantly improves when adding 10 features proposed by the expert knowledge; this unusual behavior may be explained by considering that it is the smallest selected feature set, and it may not be enough discriminative information. This behavior is schematized in Figure 8.

Concerning the selected subsets, we can note that from the 12 domain expert-suggested features, BCL2 appears in the top 10 of 3 feature selection techniques, BCLXL1 also appears in the top 10 of 3 columns in Table 7, the same occurs for XIAP and AKT. In addition, inside Table 7's top 10, BAX* and PI3K* appear in 4 columns. On the contrary, the domain expert-suggested features do not appear to impact the second half of the top 20; nevertheless, mTOR-RICTOR* and SHP2 appear in all the selected subsets, while 14-3-3*, GRB2, SOS, CytRKJAK*, and SHC are selected by 4 techniques.

### *Strengths and weaknesses*

The data generated by the Cellulat bioinformatics framework includes inputs and their discrete labels; this presents the opportunity to tackle the supervised learning task as a classification problem. Random forest (RF), support vector machines (SVM), and ANNs are some of the most prevalent classification algorithms. A common idea in big data and machine learning (ML) is that the more data you have, the more accurate your results will be; nevertheless, massive datasets come with their own set of problems. Unstructured data formats,

**Table 7.** First 20 features were obtained after applying several feature selection techniques.

| PREDICTOR'S RELATIVE IMPORTANCE | LOW VARIANCE | CHI$^2$ | LASSO | EXTRA TREES | PEARSON CORRELATION |
|---|---|---|---|---|---|
| 1 | BAX* | BCLXL1 | XIAP | E1F4E | TSC2* |
| 2 | AMPK* | PI3K* | BCLXL1 | BCLXL1 | XIAP |
| 3 | RAS* | 4EBP1 | BAX* | 14-3-3* | FOXO* |
| 4 | PI3K* | BCL2 | mTOR-RICTOR* | AKT | FKHR* |
| 5 | BCL2 | PDK1* | mTOR1* | STAT* | CRAF |
| 6 | ULK1* | ULK1* | PI3K* | FOXO* | 14-3-3* |
| 7 | 14-3-3-BAD | SHC | RHEB | BAX* | mTOR-RICTOR* |
| 8 | PDK1* | GRB2 | BCL2 | XIAP | AKT |
| 9 | 4EBP1 | SOS | CytRKJAK* | mTOR-RAPTOR | mTOR-RAPTOR |
| 10 | AKT | CytRKJAK* | BCLXL1* | RHEB* | BAX* |
| 11 | mTOR-RICTOR* | SHP2 | AMPK* | mTOR-RICTOR* | AMPK* |
| 12 | 14-3-3* | AMPK* | GRB2 | FKHR* | RAS* |
| 13 | GRB2 | RAS* | LKB1* | P21 | SHC |
| 14 | SHP2 | AKT* | 14-3-3-BAD | CRAF | CytRKJAK* |
| 15 | SOS | BAX* | SHC | SHP2 | SOS |
| 16 | CytRKJAK* | 14-3-3 | ULK1* | CASP2* | GRB2 |
| 17 | SHC | mTOR-RAPTOR | SHP2 | LKB1* | SHP2 |
| 18 | mTOR1* | 14-3-3* | SOS | AKT* | RHEB* |
| 19 | 4EBP1* | AKT | | TSC2* | PIP3 |
| 20 | RHEB* | mTOR-RICTOR* | | PIP3 | STAT* |
| . . . | . . . | | | | |

These features were ranked by their discriminative power and, to ease its understanding, partitioned in sets of 10. This partitioning lets us analyze the shared features among techniques.

**Table 8.** Accuracy results of MLP machine learning algorithm.

| | SELECTION TECHNIQUE | CARDINALITY | ACCURACY OF | | CARDINALITY |
|---|---|---|---|---|---|
| | | | SELECTED FEATURES | SELECTED AND DOMAIN KNOWLEDGE FEATURES | |
| Feature selection techniques | Low variance | 12 | 82.71 | 92.45 | 22 |
| | Lasso | 18 | 94.85 | 94.76 | 27 |
| | Recursive feature elimination | 23 | **96.62** | **96.41** | 31 |
| | Pearson correlation | 48 | 96.54 | 95.65 | 49 |
| | Chi$^2$ | 54 | 95.84 | 95.84 | 54 |
| | Extra trees | 54 | 95.84 | 95.84 | 54 |
| | *Base line (all features)* | *77* | *96.15* | | |
| | *Domain knowledge features* | *12* | *93.64* | | |

Abbreviation: MLP, multilayer perceptron.
The bold-evidenced accuracies compare the best results of the algorithmically selected features with the algorithmically selected features plus the domain-knowledge suggested features. The italicized rows show the baseline accuracy results after using the complete set of features and the domain-knowledge suggested features, i.e., the results without feature selection and without automatic feature selection, respectively.

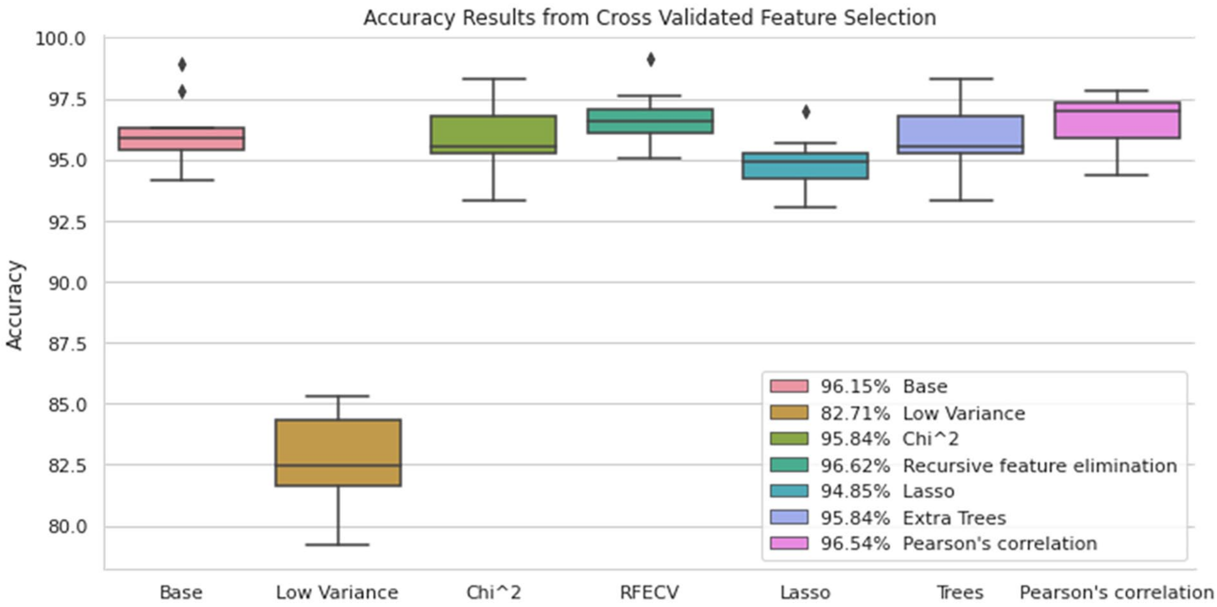

**Figure 8.** MLP accuracy results from cross-validated feature selection. Here, we can see the effect of the number of selected features by each feature selection technique on the accuracy rate; Baseline—77 features, Low Variance—12 features, Chi$^2$—54 features, Recursive Feature Elimination—23 features, Lasso—18 features, Extra Trees—54 features, and Pearson Correlation—48 features. MLP indicates multilayer perceptron.

noisy and poor-quality data, unbalanced input data distribution, unlabeled data, and high dimensionality are common problems. Another important consideration is that some machine learning algorithms were created assuming that the complete dataset could fit in memory. Big data ignores these assumptions, rendering standard algorithms useless or severely slowing them down.

Support vector machines, which try to find the optimal hyperplane with the highest margin between classes, a random forest, which can be described as an ensemble of classification trees, where each tree votes on the class assigned to a given sample, and ANNs, which can be described as parallel computing units that can separate nonlinear data, are some of the most common algorithms for supervised classification.[41] As the number of samples and classes grows, so does the complexity of these algorithms. Building a random forest, eg, becomes more expensive as the number of trees increases. In SVMs, the worst-case scenario is if the training set contains as many support vectors as samples. Although multiclass SVMs exist, their canonical implementation requires the training of a separate SVM for each class. Selecting the appropriate architecture for a specific problem in ANNs, such as the MLP used in this work, is still an open research issue.

The data mining approach to big data, empowered by machine learning techniques presented in this work, ameliorates the concerns mentioned early by acquiring, processing, and analyzing large data volumes to reduce its complexity.

## Conclusions

With this study we try to improve the traditional approach to modeling and simulation of biological systems—specifically, cell signaling networks—by integrating big data, data mining, and machine learning techniques. As a result, new inferences and knowledge were obtained from the dataset generated from the simulated system, which allowed increasing the predictive capacity of the latter.

First, the behavior of the PI3K/AKT/mTOR signaling network was modeled, simulated, verified, and calibrated; subsequently, large volumes of data describing the behavior of the simulated biological system over time were produced by running the simulation (data farming); and finally, exploratory data analysis, feature selection techniques, and analytics processes were applied to the resulting biological dataset, obtaining new inferences and knowledge about this biological system.

The resulting dataset was obtained by farming a large volume of input-output patterns produced by in silico experiments carried out by the Cellulat bioinformatics framework. These input-output patterns represent the activation/deactivation state of the 77 elements that make up the PI3K/AKT/mTOR signaling network (input pattern) and the cellular processes associated with the configuration (output pattern). The cell signaling dataset was used as input to the machine learning process using an MLP algorithm. However, for it to be helpful, we went through a cleaning and a preprocessing stage; this process involved the statistical feature selection techniques to evaluate the saliency of the features (signaling elements in the PI3K/AKT/mTOR signaling network). The predictive model resulting from applying the MLP automated learning algorithm yielded new knowledge about the simulated system by allowing the prediction of the cellular state or states associated with a specific input pattern made up of a reduced number of signaling elements.

Finally, the results of the evaluation of the machine learning model for the different selected subsets show that the use of feature selection techniques not only improves the accuracy rate of the MLP but also improves its performance, because only 30% of the original 77 characteristics are necessary to improve the baseline.

## Acknowledgements

P.P.G.-P. would like to thank the support of Universidad Autónoma Metropolitana, Unidad Cuajimalpa. M.E.S.-G. would like to thank the support of Universidad Autónoma de la Ciudad de Mexico, Unidad Cuautepec. The authors would also like to thank the support given by Cárdenas-García and Oscar Sánchez Cortés who shared the understanding of the theoretical and experimental aspects of the PI3K/AKT/mTOR signaling pathway and its role in cancer and medical data processing, respectively.

## Author Contributions

Both authors contributed equally to conceptualization, methodology, validation, formal analysis, investigation, data curation, original draft preparation, review and editing, visualization, supervision, and project administration. Both authors have read and agreed to the published version of the manuscript.

## ORCID iDs

Máximo Eduardo Sánchez-Gutiérrez https://orcid.org/0000-0003-1101-5956
Pedro Pablo González-Pérez https://orcid.org/0000-0001-7223-9035

## Data Availability

The Cell signaling dataset that support the findings of this study are available at https://raw.githubusercontent.com/elMaxPain/files/master/CRISP/PI3KAKT_50_123_500ms.arff.

## REFERENCES

1. Tolk A. The next generation of modeling & simulation: integrating big data and deep learning. Proceedings of the Conference on Summer Computer Simulation; July 26-29, 2015; Chicago, IL. New York: ACM:1-8.
2. Jacobs A. The pathologies of big data. *Commun ACM*. 2009;52:36-44.
3. Ward JS, Barker A. Undefined by data: a survey of big data definitions. *arXiv preprint arXiv:1309.5821*; 2013.
4. Laney D. 3D data management: controlling data volume, velocity and variety. *META Group Res Note*. 2001;6:1.
5. Tolk A, Diallo SY, Padilla JJ, Herencia-Zapana H. Reference modelling in support of M&S—foundations and applications. *J Simul*. 2013;7:69-82.
6. Feldkamp N, Strassburger S. Automatic generation of route networks for microscopic traffic simulations. Proceedings of the Winter Simulation Conference 2014; December 7-10, 2014; Savannah, GA. New York: IEEE:2848-2859.
7. Machado D, Costa RS, Rocha M, Ferreira EC, Tidor B, Rocha I. Modeling formalisms in systems biology. *AMB Express*. 2011;1:45.
8. Klipp E, Liebermeister W. Mathematical modeling of intracellular signaling pathways. *BMC Neuroscience*. 2006;7:S10.
9. Versari C, Busi N. Efficient stochastic simulation of biological systems with multiple variable volumes. *Electr Notes Theor Comput Sci*. 2008;194:165-180.
10. Ciocchetta F, Duguid A, Guerriero ML. A compartmental model of the cAMP/PKA/MAPK pathway in Bio-PEPA. *arXiv preprint arXiv:0911.4984*; 2009.
11. Dematté L, Priami C, Romanel A, Soyer O. Evolving BlenX programs to simulate the evolution of biological networks. *Theor Comput Sci*. 2008;408:83-96.
12. González-Pérez PP, Omicini A, Sbaraglia M. A biochemically inspired coordination-based model for simulating intracellular signalling pathways. *J Simul*. 2013;7:216-226.
13. Wurthner JU, Mukhopadhyay AK, Peimann C-J. A cellular automaton model of cellular signal transduction. *Comput Biol Med*. 2000;30:1-21.
14. Gilbert D, Fuss H, Gu X, et al. Computational methodologies for modelling, analysis and simulation of signalling networks. *Brief Bioinform*. 2006;7:339-353.
15. Ruths D, Muller M, Tseng J-T, Nakhleh L, Ram PT. The signaling petri net-based simulator: a non-parametric strategy for characterizing the dynamics of cell-specific signaling networks. *PLoS Comput Biol*. 2008;4:e1000005.
16. Hardy S, Robillard PN. Petri net-based method for the analysis of the dynamics of signal propagation in signaling pathways. *Bioinformatics*. 2008;24:209-217.
17. Albert R, Wang RS. Discrete dynamic modeling of cellular signaling networks. *Methods Enzymol*. 2009;467:281-306.
18. Morris MK, Saez-Rodriguez J, Sorger PK, Lauffenburger DA. Logic-based models for the analysis of cell signaling networks. *Biochemistry*. 2010;49:3216-3224.
19. Danos V, Feret J, Fontana W, Harmer R, Krivine J. Rule-based modelling of cellular signaling. Proceedings of the 18th International Conference on Concurrency Theory; September 3-8, 2007; Lisbon. New York: Springer:17-41.
20. Sekar JAP, Faeder JR. Rule-based modeling of signal transduction: a primer. *Methods Mol Biol*. 2012;880:139-218.
21. Rodin V, Querrec G, Ballet P, et al. Multi-agents system to model cell signalling by using fuzzy cognitive maps. Application to computer simulation of multiple myeloma. 2009 Ninth IEEE International Conference on Bioinformatics and BioEngineering; June 22-24, 2009; Taichung. New York: IEEE:236-241.
22. Reynolds ER, Himmelwright R, Sanginiti C, Pfaffmann JO. An agent-based model of the Notch signaling pathway elucidates three levels of complexity in the determination of developmental patterning. *BMC Syst Biol*. 2019;13:7.
23. Song M, Li D, Makaryan SZ, Finley SD. Quantitative modeling to understand cell signaling in the tumor microenvironment. *Curr Opin Syst Biol*. 2021;27:100345.
24. Jacques MA, Dobrzyński M, Gagliardi PA, Sznitman R, Pertz O. CODEX, a neural network approach to explore signaling dynamics landscapes. *Mol Syst Biol*. 2021;17:e10026.
25. Suthakar U, Magnoni L, Smith DR, Khan A, Andreeva J. An efficient strategy for the collection and storage of large volumes of data for computation. *J Big Data*. 2016;3:21.
26. Downward J. Targeting RAS signalling pathways in cancer therapy. *Nat Rev Cancer*. 2013;3:11-22.
27. Goodsell DS. The molecular perspective: the ras oncogene. *Oncologist*. 1999;4:263-264.
28. Neves SR, Ram PT, Iyengar R. G protein pathways. *Science*. 2022;296:1636-1639.
29. Lien EC, Dibble CC, Toker A. PI3K signaling in cancer: beyond AKT. *Curr Opin Cell Biol*. 2017;45:62-71.
30. Cárdenas-García M, González-Pérez PP, Montagna S, Cortés OS, Caballero EH. Modeling intercellular communication as a survival strategy of cancer cells: an in-silico approach on a flexible bioinformatics framework. *Bioinform Biol Insights*. 2016;10:5-18.
31. González-Pérez PP, Cárdenas-García M. Inspecting the role of PI3K/AKT signaling pathway in cancer development using an in silico modeling and simulation approach. International Conference on Bioinformatics and Biomedical Engineering; April 25-27, 2018; Granada. New York: Springer:83-95.
32. Gelernter D. Generative communication in Linda. *ACM TOPLAS*. 1985;7:80-112.
33. Rossi D, Cabri G, Denti E. Tuple-based technologies for coordination. In: Omicini, A, Zambonelli, F, Klusch, M, Tolksdorf, R, eds. *Coordination of Internet Agents*. Springer; 2001:83-109.
34. Gillespie DT. Exact stochastic simulation of coupled chemical reactions. *J Phys Chem*. 1977;81:2340-2361.
35. Gillespie DT. Stochastic simulation of chemical kinetics. *Annu Rev Phys Chem*. 2007;58:35-55.
36. Mohammed R, Rawashdeh J, Abdullah M. Machine learning with oversampling and undersampling techniques: overview study and experimental results in 2020. 11th International Conference on Information and Communication Systems (ICICS); April 7-9, 2020; Irbid, Jordan. New York: IEEE:243-248.
37. Qian Y, Liang Y, Li M, Feng G, Shi X. A resampling ensemble algorithm for classification of imbalance problems. *Neurocomputing*. 2014;143:57-67.
38. Zheng A, Casari A. *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. Sebastopol, CA: O'Reilly Media, Inc.; 2018.
39. Ramchoun H, Idrissi J, Amine M, Ghanou Y, Ettaouil M. Multilayer perceptron: architecture optimization and training. *Int J Interact Multim Artif Intell*. 2016;4:26-30.
40. Lillicrap TP, Santoro A, Marris L, Akerman CJ, Hinton G. Backpropagation and the brain. *Nat Rev Neurosci*. 2020;21:335-346.
41. Zhang C, Liu C, Zhang X, Almpanidis G. An up-to-date comparison of state-of-the-art classification algorithms. *Expert Syst Appl*. 2017;82:128-150.