big data and cognitive computing

MDPI

Article

Personalized Data Analysis Approach for Assessing Necessary Hospital Bed-Days Built on Condition Space and Hierarchical Predictor

Nataliia Melnykova [1,*], Nataliya Shakhovska [1], Volodymyr Melnykov [2], Kateryna Melnykova [3] and Khrystyna Lishchuk-Yakymovych [4]

[1] Department of Artificial Intelligence, Lviv Polytechnic National University, 79013 Lviv, Ukraine; nataliya.b.shakhovska@lpnu.ua
[2] Department of Transplantology Surgery, Danylo Halytsky Lviv National Medical University, 79010 Lviv, Ukraine; v.melnikov2013@gmail.com
[3] Lviv Emergency Hospital, 79010 Lviv, Ukraine; larkatty81@gmail.com
[4] Department of Clinical Immunology and Allergology, Danylo Halytsky Lviv National Medical University, 79010 Lviv, Ukraine; kyakymovych@gmail.com
* Correspondence: melnykovanatalia@gmail.com; Tel.: +38-0663-498-824

Abstract: The paper describes the medical data personalization problem by determining the individual characteristics needed to predict the number of days a patient spends in a hospital. The mathematical problem of patient information analysis is formalized, which will help identify critical personal characteristics based on conditioned space analysis. The condition space is given in cube form as a reflection of the functional relationship of the general parameters to the studied object. The dataset consists of 51 instances, and ten parameters are processed using different clustering and regression models. Days in hospital is the target variable. A condition space cube is formed based on clustering analysis and features selection. In this manner, a hierarchical predictor based on clustering and an ensemble of weak regressors is built. The quality of the developed hierarchical predictor for Root Mean Squared Error metric is 1.47 times better than the best weak predictor (perceptron with 12 units in a single hidden layer).

Keywords: personalization; health care; medical insurance; prediction task; machine learning; hierarchical predictor; data analysis

Citation: Melnykova, N.; Shakhovska, N.; Melnykov, V.; Melnykova, K.; Lishchuk-Yakymovych, K. Personalized Data Analysis Approach for Assessing Necessary Hospital Bed-Days Built on Condition Space and Hierarchical Predictor. *Big Data Cogn. Comput.* **2021**, *5*, 37. https://doi.org/10.3390/bdcc5030037

Academic Editor: Min Chen

Received: 29 June 2021
Accepted: 11 August 2021
Published: 16 August 2021

1. Introduction

In the era of active development of digital technologies, personalized medicine is focused on the processing of heterogeneous medical data, and technology ensures this process is manageable. Personalization is becoming increasingly important in the scientific community for many reasons, including the new optimization methods and algorithms that help reduce medical costs and provide quality and effective health care. This study focuses on solving the problem of formalization of the studied object and the formalization of its conditions during treatment or rehabilitation, which will optimize treatment processes, analyze individual patient characteristics, and forecast possible personalized solutions for medical care based on patient health assessment.

The problem of personalizing data requires a clear strategy to determine the main stages of information processing, given in Figure 1.

Determining the individual characteristics needed to solve the personalization problem depends on the key factors of object identification. For the formalized representation of the studied object in medicine, the main parameters of its general condition with certain characteristics are considered.

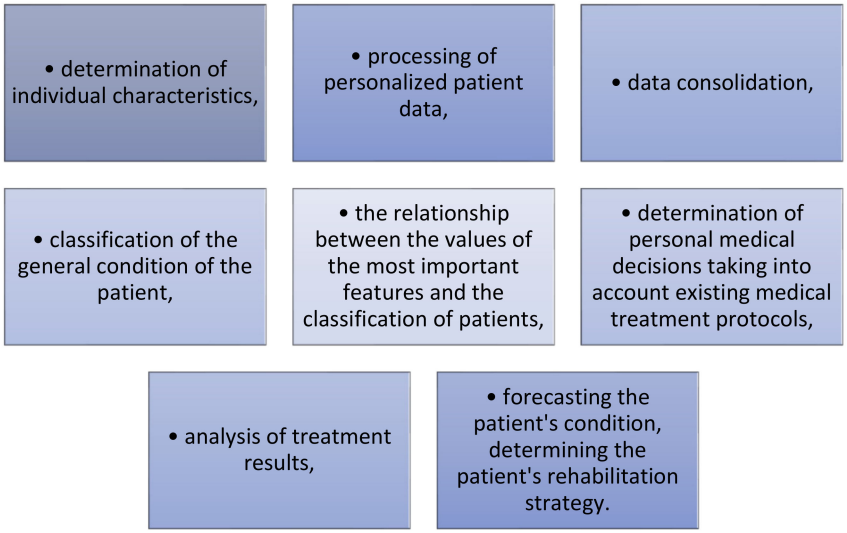

Figure 1. The main stages of patients' information processing.

Patient-oriented data are data that the patient or his/her relatives report to the doctor. The quality of patient-oriented data depends on the level of literacy regarding the patient's health, while the quality of patient data depends on the quality and sensitivity of devices and technologies involved in the collection process [1]. Patient-oriented data have several benefits, such as enabling medical data management and supporting double-checking medical research to reduce the incidence of false-positive and false-negative outcomes.

Data from the study object, coming from various sources, are used for medical and health purposes. However, data confirming the provision of special medical care are mainly concentrated and can be collected as part of screening for disease or diagnostic processes. The results of screening or diagnosis should be checked (principles of double-checking) by medical examinations. As clarified in [1], a paradigm shift in screening for disease is essential to, on the one hand, collect sufficient individualized and accurate data for patient-related medical treatment, with the aim of pre-detection, prevention, and prediction of any human health problems. On the other hand, a paradigm shift would help reduce medical costs and reduce morbidity and mortality. In addition, a paradigm shift would pave the way for personalized medicine.

Personalized medicine (PM) can be considered a continuation of traditional approaches to understanding and treating diseases with higher accuracy. The patient's gene variations profile can guide the choice of drugs or treatment protocols that minimize harmful side effects or provide more successful results. PM can also indicate a person's susceptibility to certain diseases before they manifest, allowing doctors and patients to develop a monitoring and prevention plan.

In the digital age, personalized medicine must be supported by small and big data (and should become a data-driven process) and use artificial intelligence technology to support risk prevention, prediction, and detection, and medical intervention. PM is becoming increasingly important in the scientific community (medical, paramedic, and bioengineering) for many reasons, particularly given the greater efficiency and effectiveness of health care, the reduction in medical costs, and more.

PM requires genetic information, information about predisposition to medical diseases, and accurate medical data for the patient, excluding meaningless data. PM in its real form still faces challenges and problems, such as collecting relevant and precise medical treatment data for better diagnosis and prognosis (disease screening, etc.). According to Professor Nisar Malek, Medical Director of the Clinic of Gastroenterology, Hepatology and Infectious Diseases at Tübingen University Hospital and Head of the Tübingen Center for Personalized Medicine, today's PM goals are multifaceted. He stated the following in an interview [2] that he gave in 2018: "Our first goal—and this is also our main goal—is to

further improve diagnosis and treatment for an individual patient and it is desirable to integrate this into as many medical specialties as possible. Our second goal is to collect diagnostic and treatment data and transfer them to databases, making them suitable for medical and bioinformatics processes for conclusions about a particular patient. Here we are touching on the realm of big data".

It is a well-known fact that artificial intelligence gives us many opportunities to understand and solve many complex problems in the practical and scientific space. Artificial intelligence can be used in systems designed to detect, track, and predict disease outbreaks. The better we can track the spread of a virus, the more effectively and quickly we can fight it [3].

Data mining, including in medicine, is used by several scientists. For example, Bidyuk P. in [4], for the analysis of data gaps and their filling, proposed to use decision trees, an EM algorithm, and a regression approach to forecast missing data using forecasting functions. Similar results were obtained by Sokolova O. [5] and Sina Khanmohammadi [6] for the associative classification of medical data and by Anthony Costa Constantinou for a comprehensive survey [7] and analysis of data from intelligent Bayesian models to support medical decisions. Bayesian networks are also used to reformulate the Quick Medical Reference (QMR) model in decision theory. However, with the spread of big data technology, Bayesian networks have been slow. Therefore, in the work of Y. Tang [8], Bayesian networks' parallelization was developed. Bayesian networks are also used to diagnose dementia, Alzheimer's disease, and mild cognitive impairment. The Bayesian belief network also is used in [9] and [10] for ageing diseases analysis. However, even under conditions of parallelism, for multi-parameters, large volume, and dynamic medical data, Bayesian networks should be used in combination with other machine learning methods only. An apparatus of artificial neural networks, including the use of fuzzy logic, is also actively proposed to analyze various medical data. Thus, in the work of Bodyansky E. and Perova I. [11], a system of rapid medical diagnosis based on auto-associative neuro-fuzzy memory is proposed. This system is characterized by the simplicity of both the architectural solution and its software implementation and provides for the diagnosis of patients with multiple parameters.

The next problem of personalized medical data processing is the imbalance of input data and small samples of collected data. These factors impose a number of limitations on applying existing methods and means of computational intelligence to solve such problems. In the paper [12], an approach for collecting medical data using the Internet of Things and a proposed ensemble of neural networks to detect unusual human movements is developed. However, in cases of solving highly specialized problems, which is typical for medicine, the error of learning in the ensemble of neural networks is much higher than the error of one network. In addition, the procedures for training the ensemble of neural networks are quite time- and resource-intensive. Dyvak M. [13] uses interval estimators to assess the object's condition in time constraints. In Silva-Ramírez E. L. [14], to solve classification and fill data gaps, a multilayer perceptron is trained according to different rules and an approach to multiple counting, which is based on a combination of multilayer perceptron and k-nearest neighbors. Refs [15,16] present mathematical methods for converting dynamic failure trees into Markov models, Bayesian models, or simulation models based on the Monte Carlo method, and they search for methods to reduce model complexity and increase computational performance. However, it is necessary to develop new heuristic approaches that reduce the dimensionality of the Markov model and the number of operations required for their automated analysis and calculation.

The application of Markov models can be used to calculate the parameter of the failure rate and the probability estimation of the first and second kinds of errors, and the analysis of the causes of system failure remains a topical issue [17]. Thus, data mining techniques are used to solve many problems in the processing and analyzing of medical information. However, there are no comprehensive studies aimed at identifying the patient's condition without specifying history. Some issues have been solved towards this end. Still, the

researchers only partially consider the phenomena of big data and the Internet of Things, in-depth analysis, and visualization of accumulated data to support decision-making on personalized treatment [18].

The mathematical apparatus of Petri net and its modifications have been used by researchers to model processes of different natures. For example, to model real-time planning processes in a resource-limited environment, Italian scientists in [3] proposed to use an apparatus of color Petri nets (TCPN) [19], which allowed for analysis of the interdependence, conflicting priorities, and variability of available resources (for example, in industrial projects). Spanish researchers [20] used the Petri color grid apparatus to improve the process of modeling, analysis, and semantic validation of complex system events (for example, processing technology and large data streams in the process of determining the level of air quality).

This paper describes the problem of data personalization by determining the individual characteristics needed to solve the personalization problem in the first section. The next section analyzes patient information to help identify general individual characteristics and present the mathematical model of the condition space in cube form as a reflection of the functional relationship of the general state to the studied object. The Results section presents a hierarchical predictor for a patient's number of days in hospital based on individual parameters. The Discussion section describes the main point of researching the problem of personalization and solving tasks of estimation correlation between the individual characteristics, and it evaluates the effectiveness of the chosen prediction model.

The main contributions of this paper are the following:

- The mathematical formalization of the condition space for personalized medical data is developed. It allows for predicting target variables in a sub-part of the space with higher predictive accuracy.
- A dataset with information about personal parameters of 51 patients was collected, which allowed generalization and deeper analysis (https://doi.org/10.6084/m9.figshare.14865411.v1, accessed on 28 June 2021).
- The hierarchical predictor consists of splitting objects and prediction for each separated cluster is developed. It produces 1.47 times greater predictive accuracy than the best weak predictor (perceptron with 12 units in single hidden layer).
- The collected dataset is too small. That is why a specific method based on the hierarchical predictor is proposed for small dataset analysis. Five-fold cross-validation is also used for results validation.

2. Materials and Methods

To objectively evaluate the object under study, it is necessary to build a formal model. Personalized medical data are considered as a set, the elements of which are the parameters of the object's conditions, namely the elements of sets of time-independent characteristics (A_{in}) and time-dependent parameters of the object (A_t).

Using the theory of functional analysis, time-dependent data can be formalized as a set A_t, the elements of which are subsets of individual patient's parameters $A_1, A_2, \ldots, A_n$:

$$A_t = \{A_1, A_2, \ldots, A_n\} \quad (1)$$

where:

$$A_1 = \{a_{11}, a_{12}, \ldots, a_{1m}\},$$
$$A_2 = \{a_{21}, a_{22}, \ldots, a_{2m}\},$$
$$A_n = \{a_{n1}, a_{n2}, \ldots, a_{nm}\},$$

namely:

$$A_t = \{(a_1, a_2, \ldots, a_n) | a_1 \in A_1, a_2 \in A_2 \ldots, a_n \in A_n\}$$

The system for assessing the condition of the object to optimize the recovery process is presented as:

$$Fe = A_{in}, A_t, K, ES, S, G, R, \quad (2)$$

where Fe is a system for estimating the object's conditions; A_{in} is a set of time-independent parameters, characterizing the indicators, obtained from the dynamics changes of the conditions, calculated based on the medical historical data; K is the set of coefficients, indicating changes in the performance indicators A_{in}; S is the set of strategic decisions to assess the state of the studied object; R represents the production rules for decision-making for finding optimal conditions; ES is a set of estimated conditions of the studied object, which depends on the evaluation coefficients; A_t is the set of characteristics of the studied object, which change under the influence of time; and G is medical treatment protocol, which depends on A_{in}.

At the stage of analysis of the results of researching the patient's condition, we can formulate the following property: time-dependent data at a certain point in time take constant indicators, which determine treatment decisions under the influence of personalized data.

A patient's data are characterized by heterogeneity, which complicates the analyzing process, as there is a need to formalize the patient's physical condition (FS), taking into account time-dependent (A_t) and time-independent data (A_{in}). To do this, we can represent the formalization of the patient's physical condition as a reflection of the physical condition of the object:

$$FS(t) : A_{in} \rightarrow A_t \quad (3)$$

where $A_t = A_t(t)$.

Based on the personalized patient data obtained during the process of caring for the condition and beyond, we can simulate the condition space of the studied object in time as a Euclidean space, where each pair of elements a_1, a_2 corresponds to a real number (a_1, a_2) that satisfies the conditions (axioms of the scalar product):

$$(a_1, a_1) \geq 0, \text{ where } (a_1, a_1) = 0, \text{ when } a_1 = 0$$

$$(a_1, a_2) = (a_2, a_1)$$

$$(a_{11} + a_{12}a_2) = (a_{11}, a_2) + (a_{12}, a_2)$$

$$(\lambda a_1, a_2) = \lambda(a_1, a_2)$$

Given that the condition space is represented as Euclidean space, it is possible to develop a mathematical model of condition space as a multidimensional system in the time domain.

The condition space is presented as a three-dimensional cube, where the x-axis reflects the time $Po(t)$, the y-axis is the parametric $Po(a)$, and the z-axis is an indicator for individual cases, i.e., the set of studied objects Ob (Figure 2).

As determined by simultaneous iteration, the functional relationship of the physical state to the personalized medical data reveals a relationship between the time factor and the parametric index:

$$FS_o(t) : \ P_o(t) \rightarrow P_o(a) \quad (4)$$

$$FS_o(t) = FS_o(A_{in}, A_t(t)). \quad (5)$$

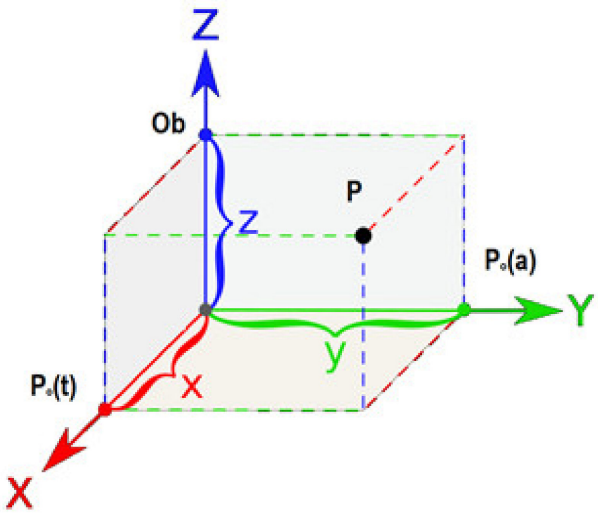

Figure 2. Condition space model of the personalized medical data.

Under the condition of the analysis of a physical condition of the investigated object at the following time iteration, dependence of time indicators of the investigated object on previous values of its physical condition is observed.

$$A_{t+1}(t) = FS_o(A_{in}, A_t(t)) \tag{6}$$

$$FS_{o+1}(t) = FS_{o+1}(A_{in}, A_{t+1}(t))$$

$$A_{t+n}(t) = FS_o(A_{in}, A_{t+n-1}(t))$$

We can represent the physical state of K at each time iteration as a relation in the functional form of the record:

$$FS_{oi} \subseteq (A_{ini}, A_t(t)_{i1}) \otimes (A_{ini}, A_t(t)_{i2}) \otimes (A_{ini}, A_t(t)_{i3}) \ldots \otimes (A_{ini}, A_t(t)_{in}). \tag{7}$$

Figure 3 represents the model of behavior of the physical state of the object.

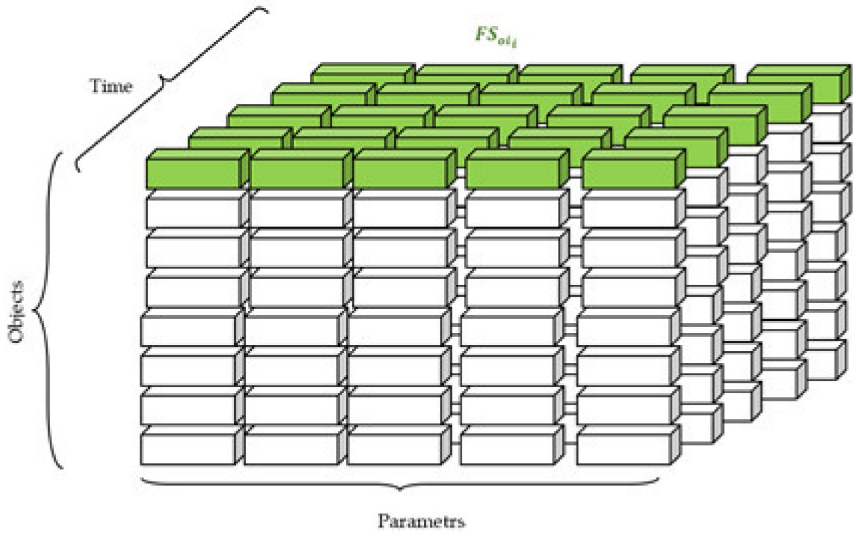

Figure 3. Model of behavior of the physical state of the object.

FS_{oi} is only an indicator of the physical condition of SO during the process of treatment or rehabilitation. Due to this, the presented relationship can be analyzed to understand the behavior of the object under study.

An important factor is the analysis and application of protocols according to the identified disease. Treatment protocols offer clear guidelines for practitioners and improve the quality of clinical decisions, including among those professionals who are accustomed to outdated medical practice. Another important advantage of clinical protocols is that they promote coherence in patient care provision at all levels.

Therefore, the condition space of the object should take into account patient's conditions at the appropriate stage of treatment and the medical protocol:

$$PFS_{oi} \subseteq FS_{oi1}(t) \otimes FS_{oi2}(t) \otimes \ldots \otimes FS_{oin}(t) \quad (8)$$

$$PFS_{oi} \subseteq FS_{oi1}(A_{ini}, A_t(t)_{i1}) \otimes FS_{oi2}(A_{ini}, A_t(t)_{i2}) \otimes \ldots \otimes FS_{oin}(A_{ini}, A_t(t)_{in}) \quad (9)$$

$$PFS_o(t) = PFS_o(FS_o(t))$$

$$PFS_{o+1}(t) = PFS_{o+1}(FS_{o+1}(t))$$

It is a well-known fact that many patients are diagnosed with a number of additional comorbidities that need to be addressed. Each concomitant disease requires treatment according to current protocols. Thus, it is necessary to take this into account when modeling the condition space of a particular object (Figure 4).

$$RPFS_{oi} \subseteq PFS_{oi1}(t) \otimes PFS_{oi2}(t) \otimes \ldots \otimes PFS_{oim}(t) \quad (10)$$

$$RPFS_{oi} \subseteq PFS_{oi1}(FS_{o1}(\mathrm{t})) \otimes PFS_{oi2}(FS_{o2}(\mathrm{t})) \otimes \ldots \otimes PFS_{oim}(FS_{on}(\mathrm{t})) \quad (11)$$

$$RPFS_o(t) = RPFS_o(PFS_o(t)) \quad (12)$$

$$RPFS_{o+1}(t) = RPFS_{o+1}(PFS_{o+1}(t)) \quad (13)$$

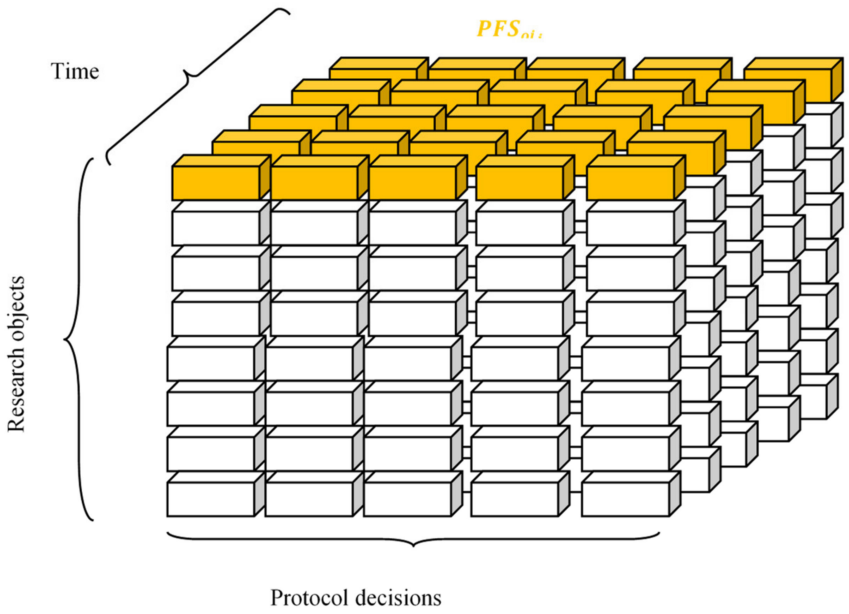

Figure 4. Model of the condition space of the object, taking into account the medical protocol.

Therefore, the state of the studied object GS_o is the result of the application of the operation of intersection between all the state spaces of the studied object at points in time $RPFS_o(t)$, $PFS_o(t)$, and $FS_o(t)$(Figure 5):

$$GS_{oi} = RPFS_o(t) \cap PFS_o(t) \cap FS_o(t) \quad (14)$$

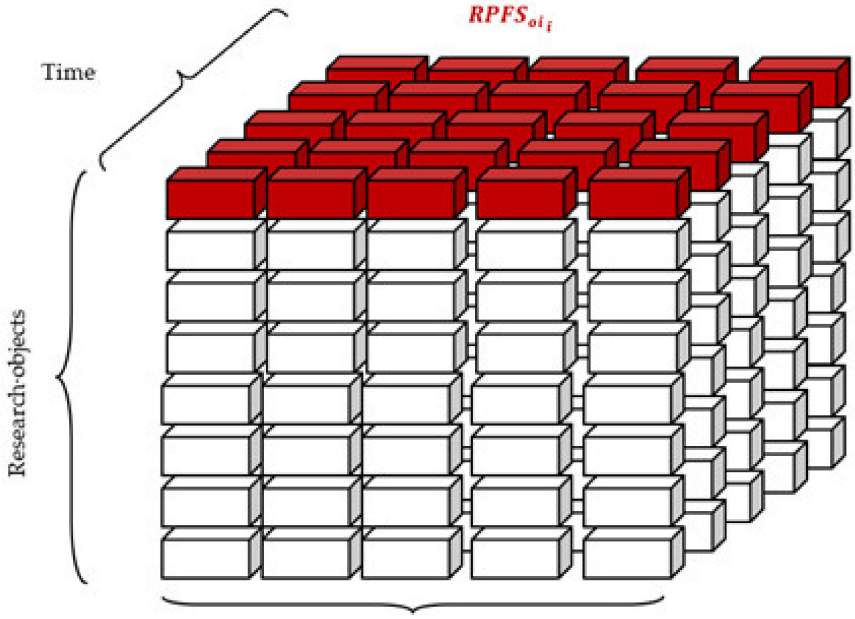

Figure 5. The condition space based on protocol and related diseases.

3. Results

3.1. The Experimental Setup

The experimental setup is organized as follows:

- Exploratory data analysis (feature normalization and encoding);
- The condition space development;
- Weak predictors selection;
- The hierarchical predictor development;
- Results evaluation.

All calculations were made using RStudio. Data were passed through Data Sampler for balancing. As results, all instances are taken into account. This means that the collected dataset is balanced.

3.2. Dataset Description

We processed patients' personalized data from collected dataset (https://doi.org/10.6084/m9.figshare.14865411.v1, accessed on 28 June 2021). This dataset was collected in the surgical department at Lviv Public Hospital (Ukraine). Patients were treated clinically for postoperative complications in the abdomen.

The dataset consists of the following characteristics:

- Age (time-dependent parameter, performance indicator A_{in})—integer;
- Sex (time-independed parameter)—boolean;
- Weight (time-dependent parameter, performance indicator A_{in})—categorical;
- Date admission—date;
- Diagnosis (time-independent parameter, used for choosing protocol PFS_{oi})—categorical;
- Related diagnosis (time-dependent parameter, $RPFS_{oi}$)—categorical;
- Flora (time-dependent parameter, $RPFS_{oi}$)—categorical;
- Medicament (time-dependent parameter, depends on PFS_{oi})—categorical;
- Active substance (time-dependent parameter, depends on PFS_{oi})—categorical;
- Time in hospital (time-dependent parameter, bed-days in hospital, target parameter)—integer.

Each instance represents the object $GS_{o.}$

The task is to predict patients' number of days in the hospital (the duration of treatment) based on drug treatment and patients' personal parameters.

Dataset consists of 51 instances and 10 parameters. Time in hospital is the target variable. After the preprocessing stage and one-hot-encoding usage for categorical variables, the dataset consists of 39 features. The missed values are found in the Flora feature and in the Related diagnosis feature. The missing data imputation procedure is not used, missing data are empty values.

3.3. Condition Space Development

For the condition space development, the following steps are performed:

- the most significant features selection;
- the splitting of instances into clusters with similar time-dependent and time-independent parameters.

The initial feature selection is performed by correlation matrix, Boruta, and regression tree. Hard voting is used for final feature selection.

The correlation between parameters is given in Figure 6.

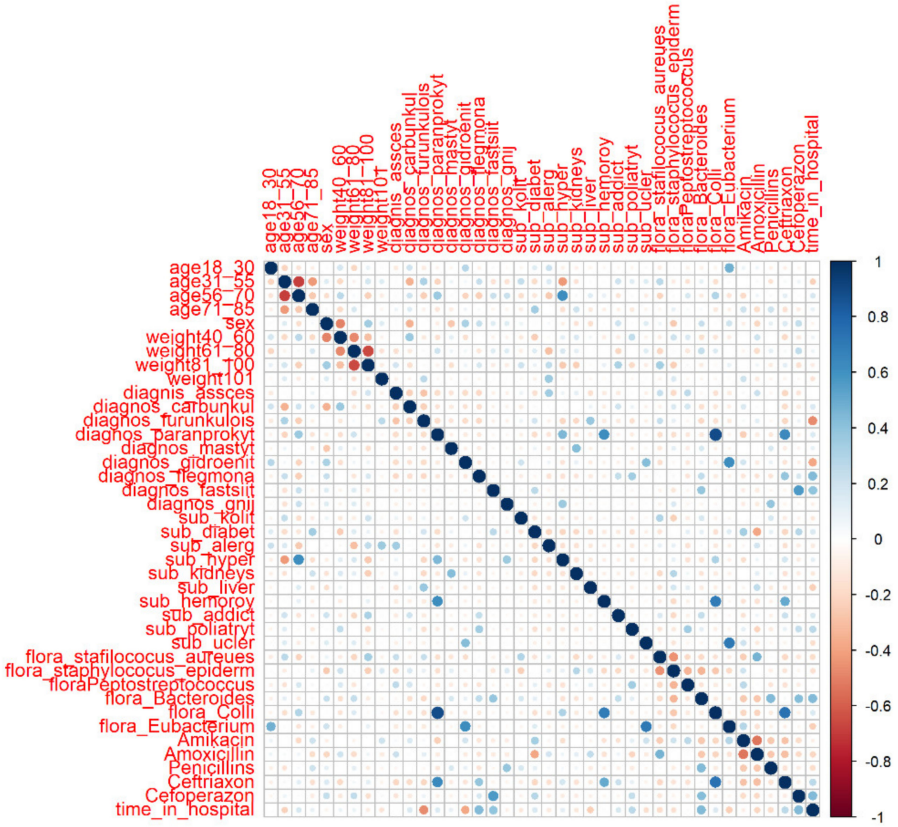

Figure 6. Correlation matrix.

Significant correlation is absent.

Next, the Boruta algorithm [21] is used (Figure 7).

The Boruta algorithm is a wrapper built around the random forest classification algorithm. Figure 7 shows "important" in green and "tentative" in yellow. Blue boxplots correspond to minimal, average, and maximum Z score of a shadow attribute.

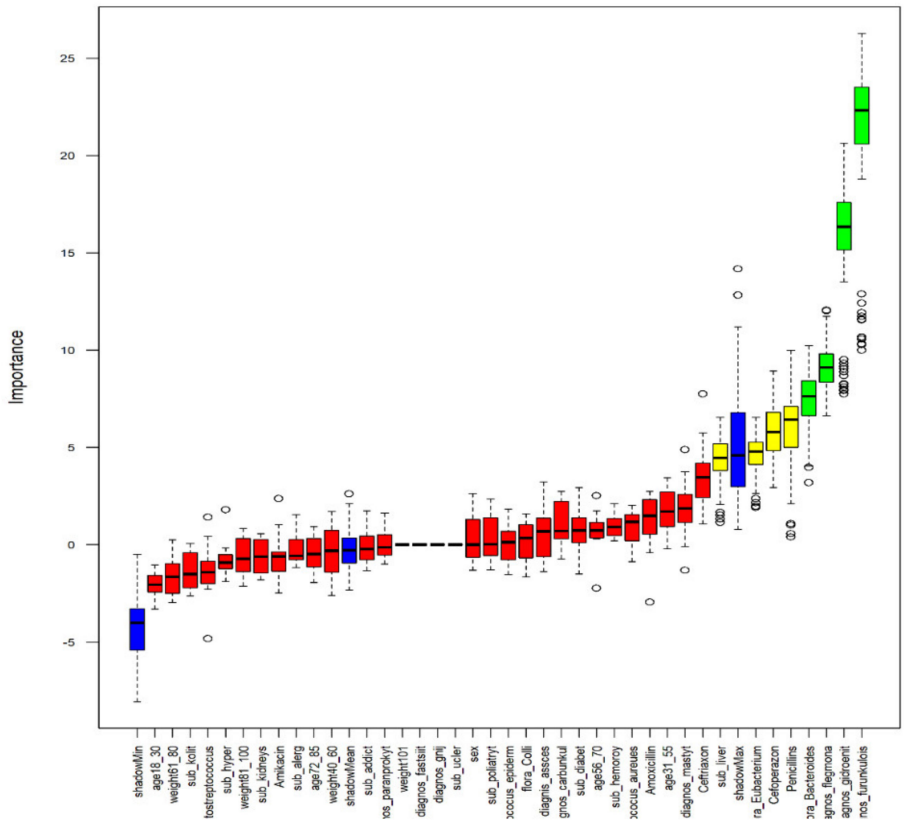

Figure 7. Result of Boruta algorithm.

The important variables are given below.

	meanImp	decision
diagnos_furunkulois	18.631571	Confirmed
diagnos_gidroenit	9.901603	Confirmed
diagnos_flegmona	8.915723	Confirmed
sub_diabet	7.557117	Confirmed

Next, a regression tree is used for features selection.

The regression tree looks as follows:

1) root 51 1981.92200 10.372550
2) diagnos_furunkulois>=0.5 11 17.63636 4.818182 *
3) diagnos_furunkulois< 0.5 40 1531.60000 11.900000
6) sub_diabet< 0.5 33 1075.87900 11.060610
12) Ceftriaxon< 0.5 26 818.65380 10.115380
24) sex>=0.5 12 602.66670 8.666667 *
25) sex< 0.5 14 169.21430 11.357140 *
13) Ceftriaxon>=0.5 7 147.71430 14.571430 *
7) sub_diabet>=0.5 7 322.85710 15.857140 *

The given above structure of the regression tree allows us to find the significant attributes and values of these attributes. The regression tree is built on the following features:

- diagnos_furunkulois;
- sub_diabet;
- Ceftriaxon;
- sex.

We can see that important variables for both methods are similar.

The hard voting result from the three feature selectors is given below:

- diagnos_furunkulois;

- sub_diabet;
- diagnos_gidroenit;
- Ceftriaxon;
- diagnos_flegmona;
- sex.

The treatment using the drug Ceftriaxon affects the duration of stay in hospital. Therefore, only time-dependent parameters are important for predicting days in hospital.

Next, clustering is used. Visual Assessment of (Cluster) Tendency (VAT) is used for analysis of the possibility for objects splitting. VAT shows poor cluster tendency (Figure 8). Small dissimilarities are represented by dark shades, and large dissimilarities are represented by light shades [22].

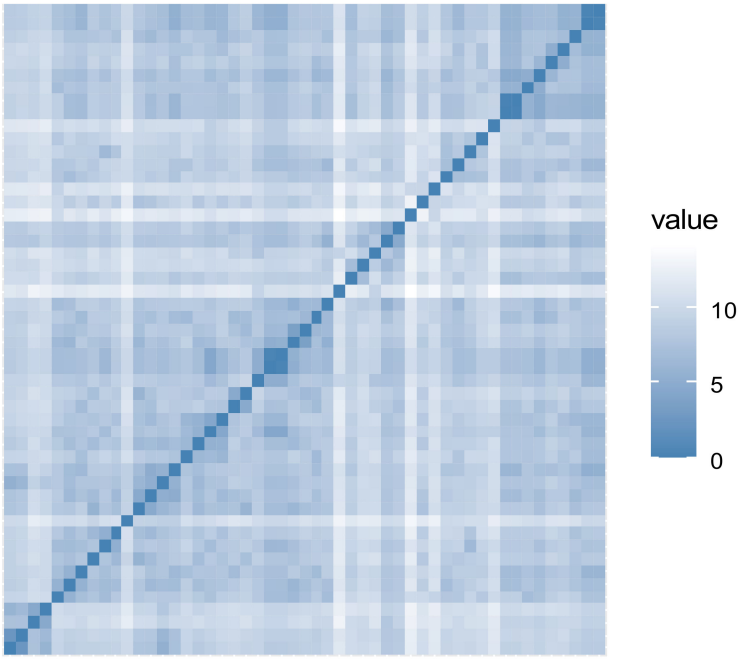

Figure 8. Results of using VAT.

The Hopkins statistic [23] is equal to 0.71. This means that dataset is not significantly clusterable. The fuzzy c-mean shows better results (Table 1). All strong instances in each cluster are marked in bold.

The same result appears with k-means visualization (Figure 9). Clusters are found to overlap, especially clusters one and three.

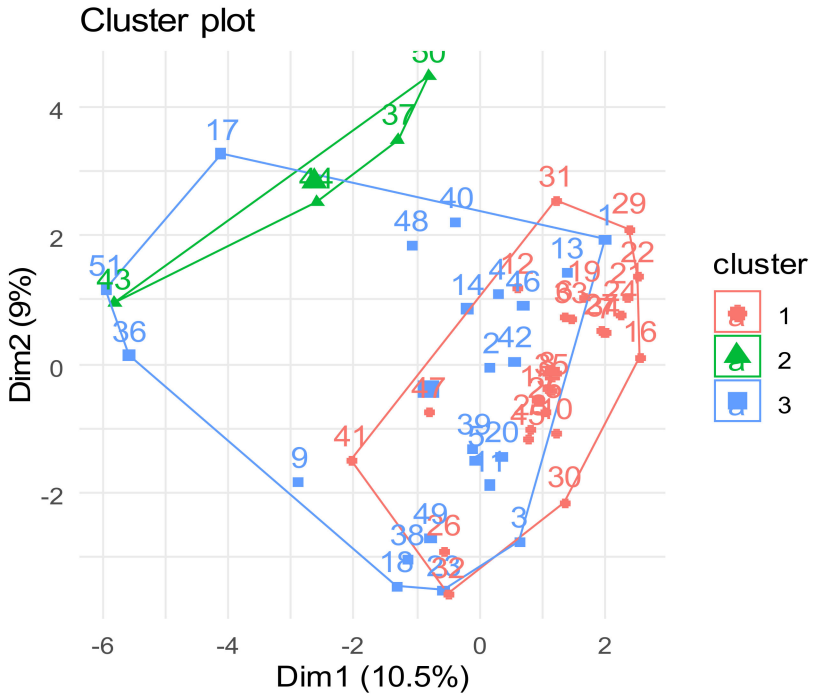

Figure 9. Results of clustering using k-means.

Table 1. Results of fuzzy c-mean.

Memberships	1	2	3	Memberships	1	2	3
[1]	**0.89150163**	0.06212372	0.046374647	[26]	0.31107702	0.65580938	0.033113601
[2]	**0.77796788**	0.08358509	0.138447034	[27]	0.15995697	**0.81908319**	0.020959847
[3]	0.56614562	0.09556459	0.338289785	[28]	0.07447833	**0.90846830**	0.017053375
[4]	**0.86018278**	0.10377855	0.036038668	[29]	0.07030249	**0.91359600**	0.016101503
[5]	0.68808007	0.09069853	0.221221402	[30]	0.05056537	**0.94033247**	0.009102165
[6]	0.03321895	**0.96077754**	0.006003508	[31]	0.07316832	**0.90999577**	0.016835911
[7]	0.30215319	0.66503031	0.032816504	[32]	0.07499601	**0.90979079**	0.015213199
[8]	0.14472035	**0.83639243**	0.018887226	[33]	0.05565317	**0.93171882**	0.012628002
[9]	**0.81184745**	0.15376100	0.034391551	[34]	0.47632097	0.48284576	0.040833272
[10]	0.04098111	**0.95165817**	0.007360716	[35]	0.03983866	**0.95303284**	0.007128497
[11]	**0.91012330**	0.05144097	0.038435725	[36]	**0.90736664**	0.06052699	0.032106373
[12]	0.14787502	**0.83223998**	0.019885001	[37]	0.08044003	0.03597534	**0.883584625**
[13]	0.64440750	0.31189553	0.043696971	[38]	**0.89430505**	0.07879489	0.026900054
[14]	**0.89189168**	0.06118803	0.046920295	[39]	0.55963468	0.09392417	0.346441148
[15]	0.05546220	**0.93451510**	0.010022703	[40]	**0.90959793**	0.05132587	0.039076209
[16]	0.04969225	**0.94011428**	0.010193473	[41]	0.18072499	**0.79517059**	0.024104421
[17]	0.64818641	0.30658830	0.045225286	[42]	0.78670219	0.08077068	0.132527121
[18]	**0.89320798**	0.07949093	0.027301094	[43]	0.09733532	0.02808194	**0.874582736**
[19]	0.06356521	**0.92325893**	0.013175858	[44]	0.06184420	0.02651523	**0.911640561**
[20]	0.89348356	0.07989919	0.026617253	[45]	0.48991432	0.46830619	0.041779495
[21]	0.05293598	**0.93743658**	0.009627441	[46]	**0.85462052**	0.06862292	0.076756557
[22]	0.08066868	**0.90078226**	0.018549056	[47]	0.31122970	0.65498914	0.033781156
[23]	**0.88746299**	0.08389872	0.028638293	[48]	**0.88266229**	0.08734761	0.029990095
[24]	0.06510596	**0.92012271**	0.014771324	[49]	**0.89659897**	0.05858958	0.044811455
[25]	0.07236656	**0.91123006**	0.016403377	[50]	0.17377591	0.04574853	0.780475563
				[51]	0.66297817	0.29258686	0.044434975

Therefore, the dataset is split into three clusters. Objects 7, 13 17, 20, 26, 34, 39, 42, 45, 47, 50, and 51 should be analyzed separately. The membership function value for these objects is less than 0.65. That is why a strong difference between objects cannot be found.

3.4. Weak Predictors Selection

At the next stage, we try to build several predictors. Two metrics are taken into account: Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE). Results are given in Table 2.

Table 2. Results of predictors.

Model	RMSE	MAPE
Linear regression	6.109753	0.4946111
Regression tree>	4.970676	0.4763212
Random forest (500 trees, mtree-3)	3.560431	0.3753441
knn	3.360431	0.3753441
SVM with radial basis kernel	3.194611	0.2923931
SVM with polynomial kernel	2.262171	0.2670496
ANN with 12 units in single hidden layer	2.0972937	0.2025539

The analysis with selected variables is given below (Table 3).

Table 3. The predictive accuracy for selected features.

Model Based on Selected Variables	RMSE	MAPE
Linear regression	3.972548	0.3240777
Regression tree	4.970676	0.4763212
Random forest (500 trees, mtree-3)	3.546447	0.2730968
knn	3.346447	0.2730968
SVM with radial basis kernel	3.095239	0.2386008
SVM with polynomial kernel	2.226906	0.1766799
ANN with 12 units in single hidden layer	2.059849	0.1513523

3.5. The Hierarchical Predictor Development

At the next stage, weak predictors are used for each separated cluster. Instances with unknown cluster are added to the fourth cluster. In total, four clusters are taken into account.

The hierarchical predictor is built using the following steps:

1. Fuzzy c-means divide objects into four clusters (Table 1);
2. Linear regression random forest, SVM with radial basis kernel, and SVM with polynomial kernel are used for each cluster separately;
3. Average voting on the obtained results is provided. Based on it, average value will be selected.

The predictive accuracy of the hierarchical predictor is presented in Table 4. The quality is worse than for the whole dataset.

Table 4. The predictive accuracy of proposed predictors.

Model Based on Whole Variables	**RMSE**	**MAPE**
Hierarchical predictor	1.401258	0.137792
Model Based on Selected Variables	**RMSE**	**MAPE**
Hierarchical predictor	1.401257	0.102961
Hierarchical predictor with repeated K-fold cross-validation, 5-fold, repeated 3 times	1.401125	0.102753

Repeated K-fold cross-validation is a technique used for small dataset validation. The advantage of this technique is the ability for parallelization. The result is presented in Table 4.

We can see that RMSE for a hierarchical predictor based on the whole dataset is not much higher than for selected features. On the other hand, the predictive accuracy for separated clusters is better than for the best weak predictor, artificial neural network (ANN) with 12 units in single hidden layer. The accuracy of the hierarchical predictor with repeated K-fold cross-validation is not much higher than the non-cross-validated method.

4. Discussion

The obtained results of researching the offered models for the collected data, in general, confirm a hypothesis about differences in predictive accuracy for the whole dataset and condition space built on clustering analysis and feature selection. However, the following two remarks must be made at once:

- The number of instances in the collected dataset is too small. A given hypothesis should be proved by a large dataset. We are working on extending the dataset. However, a similar approach was used in our previous research with other datasets [24,25] with the same approximate results;
- The authors suggest that the accuracy of the prediction will strongly depend on the number of empty values.

The hierarchical predictor is 1.01 times better than the best weak predictor (perceptron) for selected features (Tables 2 and 3). The quality of the developed hierarchical predictor for RMSE metric is 1.47 times better than the best weak predictor. The regression tree shows the same result for the whole dataset and selected variables. Linear regression shows 1.53 times better RMSE metrics for the selected features in comparison with whole dataset. The rest of the weak predictors show better results on selected features.

Feature selection is not only important for increasing accuracy. The hierarchical predictor training time for selected features is 1.2 times faster than for the whole dataset.

The proposed method is used for small datasets, and it is similar to [26]. Here, authors propose to improve RBF-based input-doubling method by introducing additional elements into the formula for calculating the output signal of the method. We increase the accuracy by 4% based on both MAE and RMSE compared to the basic method.

5. Conclusions

This paper presents feature selection and prediction of bed-days in hospital using condition space and developed hierarchical predictors. The dataset of personalized medical parameters was collected in a public hospital. This dataset is used to predict the number of days in the hospital. Patients were treated clinically for postoperative complications in the abdomen. Our study shows a low pairwise correlation between a huge subset of the parameters listed in the dataset. However, proper feature selection is needed to increase the quality of a prediction model.

Data preprocessing allows us to increase the quality of analysis. Boruta, regression tree, and correlation are used for feature selection. The results of the selection are formed based on hard voting using all feature selectors. Several clustering algorithms are used for splitting the object into separated clusters. This splitting allows for developing the condition space based on time-dependent and time-independent data and medical protocol.

A hierarchical predictor based on the combination of clustering results and four weak predictors for each cluster separately was developed in this paper. Therefore, the proposed algorithm shows a higher predictive accuracy compared to the best predictor perceptron.

Further research will include developing a hierarchical classifier built on other weak predictors and ensemble development.

Author Contributions: Conceptualization, N.S. and N.M.; methodology, N.M.; software, N.S.; validation, V.M., K.M. and K.L.-Y.; formal analysis, N.S.; investigation, N.M.; resources, K.M.; data curation, V.M.; writing—original draft preparation, N.M.; writing—review and editing, N.S.; visualization, N.S.; supervision, N.S.; project administration, N.M.; funding acquisition, N.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Ministry of Education and Science of Ukraine and National Research Foundation of Ukraine.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is available https://doi.org/10.6084/m9.figshare.14865411.v1, accessed on 28 June 2021.

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

References

1. Odeloui, A.E.; Edoh, T.O. A Context-Aware Machine-to-Machine-Enabled Pervasive Cardiac Telemetry for Personalizing Health Care Delivery. In Proceedings of the 2nd IEEE International Rural and Elderly Health Informatics Conference, Cotonou, Benin, 3–4 December 2018; pp. 1–8.
2. Djulbegovic, B.; Guyatt, G.H. Progress in evidence-based medicine: A quarter century on. *Lancet* **2017**, *390*, 415–423. [CrossRef]
3. Danhof, M.; Klein, K.; Stolk, P.; Aitken, M.; Leufkens, H. The future of drug development: The paradigm shift towards systems therapeutics. *Drug Discov. Today* **2018**, *23*, 1990–1995. [CrossRef]
4. Kuznietsova, N.V.; Bidyuk, P.I. Business intelligence techniques for missing data imputation. *Sci. News Natl. Tech. Univ. Ukr.* **2015**, *5*, 7–56.
5. Mishyna, M.; Volokh, O.; Danilova, Y.; Gerasimova, N.; Pechnikova, E.; Sokolova, O. Effects of radiation damage in studies of protein-DNA complexes by cryo-EM. *Micron* **2017**, *96*, 57–64. [CrossRef]
6. Khanmohammadi, S. An improved synchronization likelihood method for quantifying neuronal synchrony. *Comput. Biol. Med.* **2017**, *91*, 80–95. [CrossRef]
7. Perov, Y.; Graham, L.; Gourgoulias, K.; Richens, J.; Lee, C.; Baker, A.; Johri, S. Multiverse: Causal Reasoning Using Importance Sampling in Probabilistic Programming. In Proceedings of the Symposium on Advances in Approximate Bayesian Inference PMLR, Vancouver, BC, Canada, 8 December 2019; pp. 1–36.
8. Tang, Y.; Wang, Y.; Cooper, K.M.; Li, L. Towards big data Bayesian Network Learning-an Ensemble Learning Based Approach. In Proceedings of the IEEE International Congress on Big Data, Anchorage, AK, USA, 27 June–2 July 2014; pp. 355–357.
9. Lakho, S.; Jalbani, A.H.; Vighio, M.S.; Memon, I.A.; Soomro, S.S.; Soomro, Q.-U.-N. Decision Support System for Hepatitis Disease Diagnosis using Bayesian Network. *Sukkur IBA J. Comput. Math. Sci.* **2017**, *1*, 11–19. [CrossRef]
10. Seixas, F.L.; Zadrozny, B.; Laks, J.; Conci, A.; Saade, D.C.M. A Bayesian network decision model for supporting the diagnosis of dementia, Alzheimer's disease and mild cognitive impairment. *Comput. Biol. Med.* **2014**, *51*, 140–158. [CrossRef]
11. Perova, I.; Bodyanskiy, Y. Fast Medical Diagnostics Using Autoassociative Neuro-Fuzzy Memory. *Int. J. Comput.* **2017**, *16*, 34–40. [CrossRef]
12. Bhatt, C.; Dey, N.; Ashour, A. *Internet of Things and Big Data Technologies for Next Generation Healthcare*; Springer: Berlin/Heidelberg, Germany, 2017.
13. Podletskaya, N.I.; Divak, M.P. Information technology for the identification of the reverse laryngeal nerve during thyroid surgery. *Meas. Comput. Technol. Technol. Process.* **2015**, *1*, 151–157.
14. Silva-Ramírez, E.L.; Pino-Mejías, R.; López-Coello, M. Single imputation with multilayer perceptron and multiple imputation combining multilayer perceptron and k-nearest neighbours for monotone patterns. *Appl. Soft Comput.* **2015**, *29*, 65–74. [CrossRef]
15. Chiacchio, F.; Aizpurua, J.I.; D'Urso, D.; Compagno, L. Coherence region of the Priority-AND gate: 5 Analytical and numerical examples. *Qual. Reliab. Eng. Int.* **2018**, *34*, 107–115. [CrossRef]
16. Tsai, C.W.; Wu, N.-K.; Huang, C.-H. A multiple-state discrete-time Markov chain model for estimating suspended sediment concentrations in open channel flow. *Appl. Math. Model.* **2016**, *40*, 10002–10019. [CrossRef]
17. Kadri, H.; Ahmed, S.B.; Collart-Dutilleul, S. Formal approach to control design of complex and dynamical systems. *Procedia Comput. Sci.* **2017**, *108*, 2512–2516. [CrossRef]
18. Masic, I.; Miokovic, M.; Muhamedagic, B. Evidence Based Medicine—New Approaches and Challenges. *Acta Inform. Medica* **2008**, *16*, 219–225. [CrossRef] [PubMed]
19. Sobrinho, A.; Perkusich, A.; Da Silva, L.D.; Cordeiro, T.; Rêgo, J.; Cunha, P. Towards medical device certification: A colored Petri Nets model of a surface electrocardiography device. In Proceedings of the IECON 2014—40th Annual Conference of the IEEE Industrial Electronics Society, Dallas, TX, USA, 29 October–1 November 2014; pp. 2645–2651.
20. Boubeta-Puig, J.; Díaz, G.; Macià, H.; Valero, V.; Ortiz, G. MEdit4CEP-CPN: An approach for complex event processing modeling by prioritized colored petri nets. *Inf. Syst.* **2019**, *81*, 267–289. [CrossRef]
21. Anand, N.; Sehgal, R.; Anand, S.; Kaushik, A. Feature selection on educational data using Boruta algorithm. *Int. J. Comput. Intell. Stud.* **2021**, *10*, 27–35. [CrossRef]
22. Pakhira, M.K. Finding Number of Clusters before Finding Clusters. *Procedia Technol.* **2012**, *4*, 27–37. [CrossRef]
23. Li, Z.; Tian, Z.; Zhou, M.; Zhang, Z.; Jin, Y. Awareness of Line-of-Sight Propagation for Indoor Localization Using Hopkins Statistic. *IEEE Sens. J.* **2018**, *18*, 3864–3874. [CrossRef]

24. Melnykova, N.; Shakhovska, N.; Greguš, M.; Melnykov, V. Using Big Data for Formalization the Patient's Personalized Data. *Procedia Comput. Sci.* **2019**, *155*, 624–629. [CrossRef]
25. Shakhovska, N.; Izonin, I.; Melnykova, N. The Hierarchical Classifier for COVID-19 Resistance Evaluation. *Data* **2021**, *6*, 6. [CrossRef]
26. Izonin, I.; Tkachenko, R.; Dronyuk, I.; Tkachenko, P.; Gregus, M.; Rashkevych, M. Predictive modeling based on small data in clinical medicine: RBF-based additive input-doubling method. *Math. Biosci. Eng.* **2021**, *18*, 2599–2613. [CrossRef] [PubMed]