computation

MDPI

Article

Cascade-Based Input-Doubling Classifier for Predicting Survival in Allogeneic Bone Marrow Transplants: Small Data Case

Ivan Izonin [1,*], Roman Tkachenko [2], Nazarii Hovdysh [1], Oleh Berezsky [3], Kyrylo Yemets [1] and Ivan Tsmots [4]

[1] Department of Artificial Intelligence, Lviv Polytechnic National University, 79013 Lviv, Ukraine; nazarii.hovdysh.knm.2020@lpnu.ua (N.H.); kyrylo.v.yemets@lpnu.ua (K.Y.)
[2] Department of Publishing Information Technologies, Lviv Polytechnic National University, 79013 Lviv, Ukraine; roman.o.tkachenko@lpnu.ua
[3] Department of Computer Engineering, West Ukrainian National University, Lvivska, 11, 46003 Ternopil, Ukraine; olber62@gmail.com
[4] Department of Automated Control Systems, Lviv Polytechnic National University, 79013 Lviv, Ukraine; ivan.h.tsmots@lpnu.ua
* Correspondence: ivan.v.izonin@lpnu.ua

Abstract: In the field of transplantology, where medical decisions are heavily dependent on complex data analysis, the challenge of small data has become increasingly prominent. Transplantology, which focuses on the transplantation of organs and tissues, requires exceptional accuracy and precision in predicting outcomes, assessing risks, and tailoring treatment plans. However, the inherent limitations of small datasets present significant obstacles. This paper introduces an advanced input-doubling classifier designed to improve survival predictions for allogeneic bone marrow transplants. The approach utilizes two artificial intelligence tools: the first Probabilistic Neural Network generates output signals that expand the independent attributes of an augmented dataset, while the second machine learning algorithm performs the final classification. This method, based on the cascading principle, facilitates the development of novel algorithms for preparing and applying the enhanced input-doubling technique to classification tasks. The proposed method was tested on a small dataset within transplantology, focusing on binary classification. Optimal parameters for the method were identified using the Dual Annealing algorithm. Comparative analysis of the improved method against several existing approaches revealed a substantial improvement in accuracy across various performance metrics, underscoring its practical benefits.

Keywords: small data approach; input-doubling method; classification task; imbalanced dataset; transplantation medicine; machine learning; data scarcity; ANN; Probabilistic Neural Network; limited data; cascading; data augmentation strategies; optimization

Academic Editors: Dmytro Chumachenko and Sergiy Yakovlev

Received: 11 February 2025
Revised: 10 March 2025
Accepted: 17 March 2025
Published: 21 March 2025

Citation: Izonin, I.; Tkachenko, R.; Hovdysh, N.; Berezsky, O.; Yemets, K.; Tsmots, I. Cascade-Based Input-Doubling Classifier for Predicting Survival in Allogeneic Bone Marrow Transplants: Small Data Case. *Computation* **2025**, *13*, 80. https://doi.org/10.3390/computation13040080

1. Introduction

Machine learning (ML) has emerged as a transformative force across various domains, and its influence in the field of transplantology is no exception. Transplantology, the medical specialty focused on organ transplantation, has historically faced numerous challenges, including organ shortage, donor–recipient matching, and post-transplant complications [1]. However, the integration of machine learning into this field promises to address these challenges with unprecedented precision and efficiency.

One of the most critical aspects of transplantology is the matching of donors and recipients. Traditionally, this process has relied on histocompatibility tests, blood type

compatibility, and a range of other medical factors. Machine learning algorithms can significantly enhance this process by analyzing complex datasets that include not only basic compatibility factors, but also intricate patterns related to patient outcomes. Advanced ML models can predict the likelihood of organ rejection, assess the long-term success of transplants, and even consider genetic factors that were previously too complex to analyze manually [2]. By utilizing these predictive capabilities, ML can help prioritize organ allocation in a manner that maximizes the success rate and minimizes waiting times for patients.

Machine learning's potential extends beyond the initial organ matching and allocation process. In pre-transplant care, ML algorithms can analyze patient histories, genetic information, and environmental factors to predict how well a patient will respond to a transplant [3]. This predictive power can guide personalized treatment plans and improve overall patient outcomes. Post-transplant, ML models can monitor patient data in real time, identifying early signs of complications or organ rejection. For instance, ML can be used to analyze patterns in biomarkers or imaging data, offering early warnings that enable timely interventions. This real-time monitoring can lead to more personalized treatment adjustments, ultimately enhancing the longevity and quality of life for transplant recipients.

The shortage of available organs for transplantation is a persistent issue. Machine learning can contribute to alleviating this problem by optimizing the use of available organs [4]. For example, ML algorithms can help identify potential donors who might otherwise be overlooked by analyzing data from various sources, such as medical records and demographic information. Additionally, ML can enhance the effectiveness of organ preservation techniques. Predictive models can help determine the optimal conditions for preserving organs, potentially extending the time frame during which they remain viable for transplantation. By improving organ preservation, ML can increase the pool of usable organs and reduce waste.

Machine learning is also accelerating research in transplantology. It can be used to analyze vast amounts of data from clinical trials, patient registries, and laboratory studies, uncovering new insights into transplant biology and immunology [2]. For instance, ML techniques can identify novel biomarkers associated with transplant success or failure, leading to the development of new diagnostic tools and therapeutic approaches. Furthermore, ML can facilitate the development of new immunosuppressive drugs by analyzing data on drug efficacy and side effects [1]. This approach can speed up the drug discovery process and lead to more effective treatments for transplant recipients.

Despite the broad array of machine learning methods and models available today [5], nearly all of them experience a significant drop in effectiveness or become unusable when dealing with small datasets. This issue is particularly pressing in the field of transplantation, where the scarcity of sufficient data hinders the application of existing artificial intelligence tools for training procedures.

One of the primary issues with small datasets is the model's ability to generalize [6]. Machine learning models, particularly complex ones like deep neural networks, require substantial amounts of data to learn robust patterns and make accurate predictions. When data are scarce, models are at risk of overfitting, where they perform well on the training data but poorly on unseen data. This is because the model may memorize the limited examples rather than learning generalizable patterns. Additionally, small datasets often lead to high variance in model performance. Variance refers to the model's sensitivity to fluctuations in the training data [7]. With limited examples, minor changes in the data can result in significant variations in model predictions, leading to unreliable results. This high variance can be problematic in real-world applications where consistent and stable predictions are crucial.

In small data scenarios, partitioning the dataset into training and validation subsets can be challenging [8]. Typically, a validation set is used to tune hyperparameters and assess model performance. However, with limited data, creating a meaningful validation set can lead to inadequate estimates of model performance and an increased risk of overfitting to the validation data. In addition, small datasets may not provide enough examples to capture the full variability of the features. This limitation can hinder the model's ability to understand and represent the underlying structure of the data. For instance, in image classification with small datasets, the model might not encounter enough variations in lighting, angles, or object types, leading to poor feature extraction and classification performance.

When dealing with small datasets, rare events or outliers are often underrepresented. For example, in medical diagnosis, rare but critical events may not appear frequently enough to be effectively learned by the model [4]. This scarcity can result in the model failing to recognize or appropriately react to these rare but significant cases, leading to suboptimal performance in identifying anomalies or outliers.

Therefore, this study aims to enhance the existing input-doubling classifier for small imbalanced data classification through the utilization of two PNNs and the application of response surface linearization principles.

The primary contributions of this research can be outlined as follows:

- We have enhanced the input-doubling classifier for predicting survival in allogeneic bone marrow transplants with limited data by employing two machine learning methods using cascading principles.
- We developed four algorithmic implementations of the enhanced cascade-based input-doubling classifier using an Artificial Neural Network-based classifier without training and existing machine learning algorithms, specifically for analyzing small datasets.
- We optimized the performance of the improved classifiers using the Dual Annealing method and demonstrated a significant increase in accuracy compared to several existing methods, based on six different performance metrics.

The remainder of the paper is organized as follows. Section 2 presents the state-of-the-arts. Section 3 presents the topology and principles of the Probabilistic Neural Network, describes the basic input-doubling method, and provides detailed steps for the preparation, training, and application of the improved input-doubling method. To facilitate understanding, block diagrams of the procedures for preparing and applying the enhanced method are included. Section 4 contains information about the dataset used for the problem, details of the experimental modeling of the improved method, and results from its optimized version. Section 5 provides a comparison with existing methods and outlines future research directions. The Section 6 of this paper summarizes the conclusions.

2. The State-of-the-Arts

As previously mentioned, the challenge of intelligent analysis of small datasets is increasingly prevalent in various medical applications [6], including transplantology. However, universal, effective, or even practical methods for solving regression or classification problems with limited training data are scarce in the scientific literature [9,10]. Consequently, many of these problems remain unresolved or await the availability of more extensive datasets [11].

One of the most effective artificial intelligence tools for addressing such challenges is the Probabilistic Neural Network (PNN) [12]. Its topology is illustrated in Figure 1.

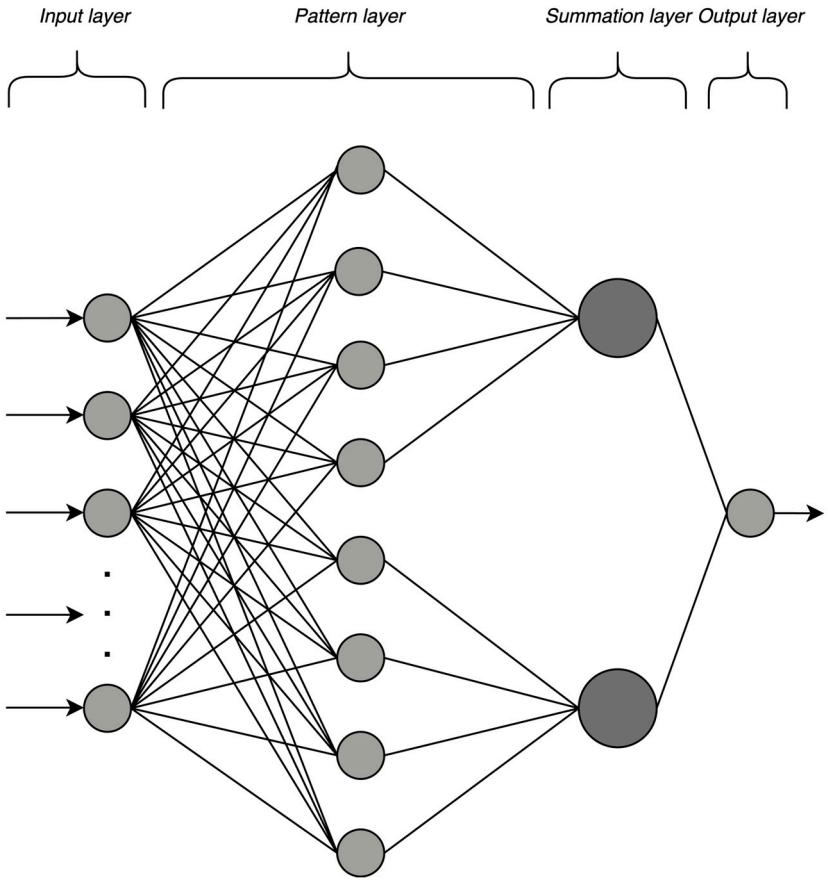

Figure 1. PNN topology for solving binary classification tasks.

This type of neural network is used to model probabilities of different classes or outcomes in classification tasks. The operation of a PNN can be described by four main steps [13]:

1. Calculating the chosen distance from the current sample to all samples in the support dataset;
2. Calculating Gaussian functions from the distances computed in the previous step;
3. Determining the probabilities of belonging to each of the predefined classes;
4. Assigning the class label with the highest probability from the set defined in the previous step.

Detailed descriptions of these steps are provided in Refs. [13,14], while this paper will focus on analyzing the numerous advantages of using PNN specifically for small data analysis.

The use of PNNs for analyzing limited datasets offers several significant advantages over traditional artificial neural networks for addressing the task at hand, including [4,13,14]:

- PNNs estimate the probability that the input data belongs to a specific class or outcome. This is useful for tasks where accounting for uncertainty or data incompleteness is important, which is very characteristic of small medical data samples, most of which are collected manually by doctors.
- PNNs often use special activation functions, such as kernel functions, to calculate probabilities for each class. One of the most well-known variants is the kernel-based network, where Gaussian functions are used to estimate probabilities. This is the type that will be used in this work.
- Probabilistic neural networks have the advantage of faster training compared to some other types of neural networks, as their training reduces to modeling probabilities

without complex error backpropagation processes, which are typical for other types of artificial neural networks.

- Probabilistic neural networks can model complex dependencies and relationships in data. They are capable of integrating additional information about uncertainty and data incompleteness, which can be important for small datasets.
- By using probabilistic approaches, this neural network can be less prone to overfitting, which is a major issue during the intelligent analysis of small datasets.
- Due to the explicit probability estimation, PNNs can be easier to interpret, as they provide information on how confident the algorithm is in its predictions.
- PNNs are characterized by high generalization properties and have only one main parameter (smooth factor), the optimal value of which should be determined experimentally for effective functioning of the artificial neural network. In this paper, we will use an optimization method to determine the value of the Gaussian function's smoothing parameter, or smooth factor (sigma).
- PNNs can use various distances between input data and reference (training) samples to assess the probability of belonging to each class. This paper will investigate the effectiveness of using a range of such functions, including Euclidean, Manhattan, cosine, Chebyshev, Minkowski, and Canberra distances.

Despite the aforementioned advantages of the PNN in analyzing small datasets and its provision of a powerful classification approach, it is not a panacea, and its effectiveness depends on the specific conditions, data, and parameters [4,14]. However, in many cases, the features described above can be very useful, especially when working with limited data.

In Ref. [15], the authors propose a new classification method for analyzing small datasets. It is based on the combined use of three main research directions in the field of small data analysis: the first is augmentation, the second is ensemble learning, and the third is nonlinear methods. Let us examine them in more detail.

Data augmentation is probably the first, most intuitive approach to improving the accuracy of a chosen machine learning method or model [8]. Classical approaches to augmentation involve using a selected method to synthesize artificial samples from both classes, thereby increasing the size of a small dataset [5]. For example, simpler methods such as columnar or row-wise methods [16] can be used, or more complex ones like Generative Adversarial Networks [17] or autoencoders [18]. However, in the case of analyzing extremely small datasets, training neural network tools to synthesize new data from such samples is quite challenging [19]. Additionally, existing methods from the data augmentation class have a number of drawbacks. The first issue with this approach is determining how many artificial samples should be generated to improve accuracy while minimally increasing computational complexity or model training time. This constraint will become an additional hyperparameter that needs to be carefully tuned in each specific case. A second, quite logical drawback of this approach is that the synthesized samples may not accurately reflect the natural state of the data being studied. Such artificial data can introduce additional noise, which, in the case of analyzing a small dataset, may lead to significant errors in model performance or even completely inadequate results.

To overcome both of these drawbacks, the authors in Ref. [15] proposed a new method for expanding the dataset. This method is based on the principle of the Cartesian product of ordered input sets. In this case, data augmentation occurs by concatenating all possible pairs of vectors within a given small training sample. The output attribute is formed as the difference between the output values of the two concatenated vectors. As a result, we obtain a dataset that is quadratically expanded, with twice the number of independent attributes. However, the most crucial aspect here is the output values on which the chosen model

will be trained. These are the differences between the outputs of the concatenated vectors. Given this, it should be noted that this augmentation procedure cannot be a standalone method, as it requires the execution of unique procedures during the implementation of the method.

Based on the augmented dataset, which in the case of using Ref. [15] is quadratically increased, it becomes possible and appropriate to apply various existing nonlinear machine learning methods. These methods will enable the consideration of nonlinearity in the dataset, which will reflect on the accuracy of the obtained results. Among such methods, both single methods (Support Vector Classifier and its modifications [20]), including ANN-based methods (PNN [13], RBF ANN [21]), and ensemble machine learning methods, including classical ones (Random Forest, Gradient Boosting [2,22,23]) and newer ones (Random RotBoost [24], Rotation Forest [25]), can be used. The approach proposed in Ref. [15] can be based on the use of any of the nonlinear and kernel-based machine learning methods or artificial neural networks described above.

When it comes to ensemble methods, they also represent a distinct class of methods that are widely used in the analysis of small datasets, known as ensemble learning [26]. Among them, four main categories of methods should be highlighted: boosting, bagging, stacking, and cascading [23,27–29]. Different authors apply various combinations of machine learning methods and artificial neural networks, using different composition rules, which in most cases demonstrate an increase in the accuracy of such models compared to each base method used in their construction. For example, the PNN can be used as a baseline algorithm (weak classifier) for more complex, particularly stacked ensemble methods of data analysis [13]. In this paper, for modifying the method from Ref. [15], we will consider the combined use of cascade ensemble and voting-based ensemble.

Voting ensembles, based on the voting principle, are also applied in many works, including in Ref. [15] during the implementation procedure of the method. However, in most cases, the implementation of such ensembles involves the use of several homogeneous (homogeneous ensembles) or heterogeneous (heterogeneous ensembles) machine learning models whose votes are aggregated [30]. The main difference here is that in Ref. [15], ensemble principles are implemented using only one machine learning method or artificial neural network. Despite all the advantages of the method from Ref. [15], which incorporates three main approaches for analyzing small datasets, in some tasks, this method may not provide the level of accuracy required for certain applications. Therefore, in this paper, we aim to address this limitation by modifying the method from Ref. [15].

3. Materials and Methods

As mentioned earlier, one promising class of methods for analyzing small and extremely small datasets is input-doubling methods [15,31]. These methods combine the benefits of data augmentation, nonlinear techniques, and ensemble learning principles. However, the data augmentation procedure in this case differs from traditional approaches, as it does not generate entirely new vectors. Instead, it expands the small dataset by leveraging the existing data. Moreover, this procedure quadratically increases the dataset size, making it suitable for applying existing machine learning (ML) or artificial neural network (ANN) methods to data analysis tasks that would otherwise be challenging with the original dataset. Additionally, the input-doubling approach is based on ensemble principles, which helps improve classification accuracy when solving the given task.

The algorithm for applying the method [15] involves the following steps: (i) concatenation of the input vector with the unknown class value with all vectors from the initial support dataset; (ii) prediction of values for the temporary matrix of vectors formed in the previous step; (iii) formation of sums from two elements—the known output value and

the corresponding prediction from the previous step; and (iv) use of the plurality voting principle to obtain the final class label for the input vector.

As shown in Ref. [15], the existing input-doubling classifier demonstrates better results than the baseline nonlinear machine learning method used in its training procedure. However, in some cases, when the data are small, high-dimensional, and complex, this method shows unsatisfactory results. This is why this paper proposes a modification of the method [15] to improve the accuracy of analysis for high-dimensional small datasets. It is based on the cascading principles [32–34].

The principle of cascading machine learning methods involves the step-by-step application of several models or algorithms, where the output of each previous stage is used as input for the next. This helps improve the accuracy and efficiency of the model, as each stage can refine or correct the results from previous ones.

One application of cascading is using the output of the first algorithm to extend the input data space, which enhances classification or prediction accuracy. This concept is supported by Cover's Theorem [35], which states that transforming data into higher dimensions can reduce the complexity of the classification task, making the models more effective.

In general, cascading allows for the use of weak models at each stage, which together can achieve much better results than a single powerful model. In this paper we apply this principle by incorporating temporary predictions as an additional attribute to the initial small dataset.

Let us examine the main steps of the modified data augmentation method (Figure 2) for training (forming the support sample for machine learning algorithm 2 (ML-2)):

1. Apply PNN to assign class labels to each vector in both the training and test samples. Note that during the classification of the training (support) dataset, the vector for which the prediction is made is temporarily removed from the sample (at the time of prediction).
2. Generate extended vectors by concatenating all possible pairs of vectors from the extended reference (training) sample. This process results in a quadratic increase in the number of vectors, with each vector having doubled input features. However, unlike the method described in Ref. [15], the dimensionality of these vectors will be increased by 2 positions due to the additional step 1.
3. Use the selected classifier (ML-2) to carry out its training procedure, if required. In this approach, the combined execution of steps 1 and 2 allows us to create a reference sample for ML-2, which can be chosen by the user.
4. As shown in Ref. [15], the existing input-doubling classifier demonstrates better results than the baseline nonlinear machine learning method used in its training procedure. However, in some cases, when the data are small, high-dimensional, and complex, this method shows unsatisfactory results. This is why this paper proposes a modification of the method [15] to improve the accuracy of analysis for high-dimensional small datasets. It is based on the cascading principles by incorporating temporary predictions as an additional step.

Figure 2 provides a visual representation of the steps described above to illustrate the modification introduced in this work. Elements highlighted in green indicate components that are not part of the original method from Ref. [15], but are newly introduced in the method discussed in this paper.

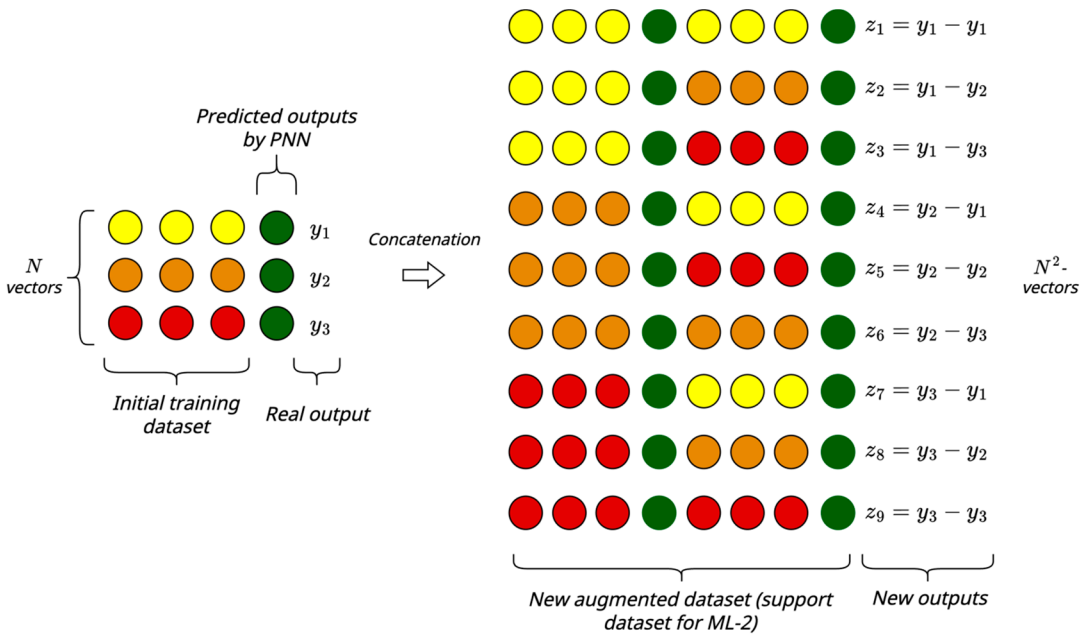

Figure 2. Modified augmentation procedure for the cascade-based input-doubling classifier.

In line with the augmentation procedure, which, as noted earlier, cannot be used as a standalone method, this paper presents an algorithm for applying the modified input-doubling technique from Ref. [15]. Figure 3 provides visualizations of the key steps involved in this algorithm:

1. Assign the class label to the current observation with an unknown outcome using the PNN and the initial dataset as the supporting dataset.
2. Create a temporary dataset for the current observation by its stepwise concatenation with each vector from the initial support (training) sample, incorporating previously predicted outcomes (highlighted in green). Unlike the method described in Ref. [15], each vector in the temporary dataset will have 2 additional dimensions due to the inclusion of step 1.
3. Apply the selected classifier (ML-2) to process this dataset and compute the differences z_i (as shown in Figure 3). In this case, we used a different Ml or ANN, as we can apply it, utilizing the quadratically augmented dataset.
4. Calculate sums from two elements—the known output value y_i and the corresponding prediction from the previous step z_i. Use the plurality voting principle to determine the final class label for the current vector with an unknown label (Figure 3).
5. Repeat steps 1–4 for all subsequent vectors with unknown class labels (from the test sample).

By leveraging cascading principles as the foundation for the proposed modification of the existing classifier, alongside custom data augmentation techniques and voting-based decision-making, the accuracy of classification tasks should be significantly improved, particularly for high-dimensional, complex, and small-sized datasets.

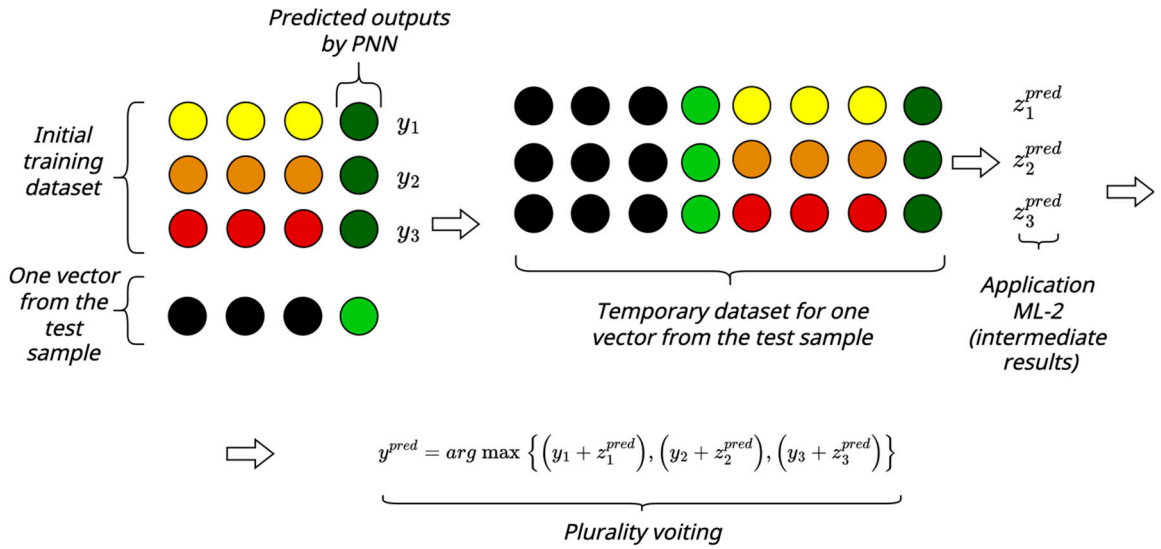

Figure 3. Modified application procedure for the cascade-based input-doubling classifier.

4. Modeling and Results

The authors have developed a custom software solution, which is a comprehensive tool for processing and analyzing medical data using machine learning methods. The software was implemented in Python (https://www.python.org/) due to its extensive capabilities in data processing and machine learning, utilizing libraries such as NumPy and Pandas for data manipulation, and Scikit-learn for implementing machine learning algorithms.

4.1. Dataset Descriptions

The modeling was carried out on a small, imbalanced dataset available in the open-access repository [36]. The dataset collection was preceded by research from Refs. [37–40]. The dataset includes pediatric patients diagnosed with various hematologic disorders, both malignant and non-malignant. Malignant conditions covered include acute lymphoblastic leukemia, acute myelogenous leukemia, chronic myelogenous leukemia, and myelodysplastic syndrome. Non-malignant cases include severe aplastic anemia, Fanconi anemia, and X-linked adrenoleukodystrophy. All patients in the dataset underwent unmanipulated allogeneic hematopoietic stem cell transplantation from unrelated donors. The dataset comprises 187 samples, each characterized by 37 attributes. All of the information about the dataset is very detailed and presented in Ref. [36]. The applied task is predicting survival in allogeneic bone marrow transplants. From a machine learning perspective, it is a binary classification task.

4.2. Performance Indicators

To assess the performance of the classification task in this study, several metrics were utilized [41,42].

Accuracy measures the overall correctness of the model, representing the proportion of instances where the class was correctly predicted:

$$Accuracy = \frac{TP + TN}{N}, \quad (1)$$

Precision measures the proportion of true positives among all instances classified as positive. It reflects the fraction of patients identified as having relapses who actually do have a relapse:

$$Precision = \frac{TP}{TP + FP}, \tag{2}$$

Recall measures the proportion of all true positive cases that the model successfully identified:

$$Recall = \frac{TP}{TP + FN}, \tag{3}$$

F1-score is the harmonic mean of Precision and Recall, which helps balance both of these metrics:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}, \tag{4}$$

Matthews Correlation Coefficient is a more comprehensive metric that takes into account all four quadrants of the confusion matrix, and is therefore a reliable metric even when analyzing results from imbalanced classes:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}, \tag{5}$$

Cohen's Kappa score helps assess the agreement between two ratings, taking into account the probability of random agreement. Here, Po is the observed agreement frequency between the ratings, and Pe is the probability of random agreement:

$$Cohen's\ Kappa = \frac{P_o - P_e}{1 - P_e}, \tag{6}$$

This paper focuses on analyzing three metrics: F1-score, Matthews Correlation Coefficient (MCC), and Cohen's Kappa score. The choice of these metrics is justified as follows. F1-score is useful because it balances Precision and Recall, both of which are critically important in medical applications. For instance, in the context of disease diagnosis, it is essential not only to identify as many true cases of the disease as possible (high Recall), but also to minimize the number of false positive diagnoses (high Precision). The F1-score helps find the optimal balance between these aspects. Matthews Correlation Coefficient (MCC) is considered one of the best metrics for evaluating the quality of binary classifiers, especially in cases of imbalanced classes, as it takes into account all four components of the confusion matrix (TP, TN, FP, FN). It provides a high score only when the model accurately classifies cases from both classes. Cohen's Kappa score assesses the degree of agreement between observed classifications and what might be expected by chance. This metric is very useful for determining how effectively the model addresses the issue of dataset imbalance, giving more weight to the accuracy of predictions in less-represented classes.

Analyzing results based on the combined consideration of these metrics allows for an objective assessment of the model's effectiveness in real-world medical diagnostics. It is crucial not only to identify as many disease cases as possible, but also to avoid false diagnoses that could lead to unnecessary treatments, undue patient anxiety, and potential legal issues related to misdiagnosis.

4.3. Cascade Method's Parameters Optimization

The enhancements discussed in this paper involve employing one PNN in the first cascade level and another ML or ANN at the second cascade level. Each method is characterized by a parameter.

In the case of utilizing a PNN, this ANN is characterized by smooth factor (sigma), which must be carefully selected [14]. In addition, the initial step in the PNN involves calculating the distance between the current observation and all vectors in the reference sample. As demonstrated in Ref. [4], the choice of distance metric significantly affects PNN performance and, by extension, the effectiveness of the improved method described here. Consequently, we investigated the impact of various distance metrics on the PNN method's accuracy. The metrics examined include Euclidean, Manhattan, Cosine, Chebyshev, Minkowski, and Canberra [4]. The results of this experiment are presented in Table A1. As shown in Table A1, the metrics F1-score, MCC, and Cohen's Kappa differ significantly depending on the chosen distance. This is why this parameter should be treated as a hyperparameter of the overall cascade method. In addition, the optimal sigma value is also treated as a hyperparameter for the entire cascade-based method.

Furthermore, any machine learning method can be used at the second level of the cascade. This possibility is explained by the quadratic increase in the data sample size for training procedures, which will be sufficient for the use of existing methods, even in the case of an initially extremely small data sample (up to 100 observations).

The parameters of the existing methods used in this work at the second level of the cascade-based input-doubling method are shown in Table 1.

Table 1. The operating parameters of the existing methods used for optimization during the implementation of the training procedure of the cascade-based input-doubling method.

Method	Parameter
Probabilistic Neural Network	{'distance type', 'smooth factor—sigma'}
RandomForest	{'n_estimators', 'max_depth'}
XGBoost	{'n_estimators', 'learning_rate', 'max_depth'}
HistGradientBoosting	{'max_iter', 'learning_rate', 'max_depth'}

4.4. Results

To achieve optimal modeling results, this study performed optimization of the cascade-based input-doubling method using the Dual Annealing optimization method [43]. This approach ensured the correct selection of all method parameters, thereby ensuring high classification accuracy.

To ensure the reliability of the modeling results, five-fold cross-validation was performed during the experiments. The results of five such experiments, using different subsets in both training and testing modes, were summarized and presented as the mean value and standard deviation for each individual metric. It should be noted that optimization was performed for each individual run, and the training time of the cascade method was taken as the total time for all five experiments (five-fold cross-validation).

For a comprehensive evaluation of the performance of the input-doubling cascade method, classification results were assessed using various metrics, including metrics (1)–(6), when solving the task of predicting survival in allogeneic bone marrow transplants using a small dataset.

Table 2 presents the results of four different algorithmic implementations of the input-doubling cascade method based on metrics (1)–(6):

- Algorithm 1—Cascade-based input-doubling method using PNN with RandomForest;
- Algorithm 2—Cascade-based input-doubling method using PNN with XGBoost;
- Algorithm 3—Cascade-based input-doubling method using PNN with HistGradientBoosting;
- Algorithm 4—Cascade-based input-doubling method using PNN-1 with PNN-2.

Table 2. Optimized results for different algorithmic implementations of the proposed cascade-based input-doubling method (mean value ± std).

Algorithm / Metric	Accuracy	Precision	Recall	F1-Score	Matthews Correlation Coefficient	Cohen's Kappa	Training Time, Seconds *
Algorithm 1 (PNN with RandomForest)	0.97± 0.021	1.00± 0.03	0.95± 0.019	0.98± 0.025	0.95± 0.022	0.95± 0.020	883.45
Algorithm 2 (PNN with XGBoost)	0.95± 0.029	1.00± 0.025	0.90± 0.032	0.95± 0.029	0.90± 0.026	0.89± 0.029	516.8
Algorithm 3 (PNN with HistGradientBoosting)	0.95± 0.022	1.00± 0.02	0.90± 0.018	0.95± 0.019	0.90± 0.019	0.89± 0.019	38,462.5
Algorithm 4 (PNN-1 with PNN-2)	0.89± 0.034	0.90± 0.036	0.90± 0.027	0.90± 0.032	0.79± 0.029	0.79± 0.030	99,100.49

* The training time calculated totally for all five-fold cross-validation procedure.

The optimal parameters for all four algorithms are presented in Table A2. It should be noted that the first cascade of the method is implemented using the PNN, as it is used in the basic input-doubling method [15].

As shown in Table 2, the highest accuracy in solving the task of predicting survival in allogeneic bone marrow transplants using a small dataset is demonstrated by Algorithmic Implementation 1, which uses a PNN at the first level of the cascade to modify the data augmentation procedure and the Random Forest algorithm, which is used to implement the training procedure of the cascade method as a whole. The lowest accuracy in this case is demonstrated by Algorithmic Implementation 4, which uses two PNNs at two levels of the cascade method.

Therefore, the first algorithmic implementation of the cascade-based method (PNN with Random Forest), presented in Table 2, was used to compare the effectiveness of the method with several existing ones.

5. Comparison and Discussion

To assess the effectiveness of the improved method, it was compared with both the existing input-doubling classifier, classical PNN, and lots of other existing machine learning algorithms and ANNs. Optimal parameters for all methods were identified using a Dual Annealing optimizer. The evaluation included metrics (1)–(6) and considered the training time. Note that the training time was recorded for the entire training sample during both optimization and five-fold cross-validation.

The findings from this comparison are summarized in Table 3. Since the dataset under analysis is imbalanced, the most appropriate metrics for evaluating the performance of the three classifiers are the F1-score, Matthews Correlation Coefficient (MCC), and Cohen's Kappa score. Execution time is also an important factor to consider.

As shown in Table 3, the KNN method demonstrates the lowest classification accuracy, with both the Matthews Correlation Coefficient (MCC) and Cohen's Kappa score falling below 0.5. However, it has the shortest execution time compared to the other methods. In contrast, other machine learning methods and artificial neural networks yield better results. For example, ensemble methods such as HistGradientBoosting and Random Forest, as well as the Multi-Layer Perceptron (MLP), achieve an analysis accuracy greater than 90% based on the F1-score for a small dataset. However, the MLP exhibits over 40 times longer training time compared to the Random Forest algorithm.

Table 3. Comparison of the different performance indicators for all investigated methods (four algorithmic implementations of the cascade-based input doubling method and existing ML algorithms and ANN).

Algorithm / Metric	Accuracy	Precision	Recall	F1-Score	Matthews Correlation Coefficient	Cohen's Kappa	Training Time, Seconds
Algorithm 1 (PNN with RandomForest)	0.97	1.00	0.95	0.98	0.95	0.95	883.45
Algorithm 2 (PNN with XGBoost)	0.95	1.00	0.90	0.95	0.90	0.89	516.8
Algorithm 3 (PNN with HistGradientBoosting)	0.95	1.00	0.90	0.95	0.90	0.89	38,462.5
Algorithm 4 (PNN-1 with PNN-2)	0.89	0.90	0.90	0.90	0.79	0.79	99,100.49
MLP	0.92	1.00	0.86	0.92	0.85	0.84	1260.97
HistGradientBoosting	0.92	1.00	0.86	0.92	0.85	0.84	43.39
RandomForest	0.92	1.00	0.86	0.92	0.85	0.84	28.51
SVM	0.89	0.95	0.86	0.90	0.79	0.79	5.565
LightGBM	0.89	1.00	0.81	0.89	0.81	0.79	41.69
XGBoost	0.89	1.00	0.81	0.89	0.81	0.79	41.42
Input-doubling method (PNN)	0.82	0.95	0.77	0.85	0.64	0.62	44,522.88
Classical PNN	0.71	0.81	0.62	0.70	0.45	0.43	6.336
KNN	0.66	0.90	0.43	0.58	0.42	0.35	1.263

When considering the cascade-based input-doubling algorithms, Algorithm 4 shows an improvement in accuracy over the baseline input-doubling method [15]. Nevertheless, existing machine learning methods, such as the Random Forest algorithm, outperform it. Additionally, due to the specific nature of the PNN procedure, it is very slow when analyzing large datasets, with performance degrading significantly as the dataset size increases quadratically. On the other hand, other algorithmic implementations, particularly Algorithms 2 and 3, demonstrate both improved accuracy on smaller datasets and more acceptable training times. Specifically, Algorithm 2 performs 86 times faster and delivers 10% higher accuracy than the baseline input-doubling method [15].

The improved cascade input-doubling classifier, Algorithm 1, designed to predict survival in allogeneic bone marrow transplants with limited data, achieves the highest performance across all metrics (1)–(9) when compared to the other methods. This enhanced method increases the F1-score by 12 points over the baseline input-doubling classifier, 28 points over the classical PNN, and 6 points over the Random Forest algorithm. The Matthews Correlation Coefficient is improved by 31 points over the baseline input-doubling method, 50 points over the PNN, and 10 points over the Random Forest. Similarly, the Cohen's Kappa score sees a significant improvement, rising by 33 points compared to the baseline, 52 points compared to the classical PNN, and 10 points compared to the Random Forest algorithm.

Despite these significant accuracy improvements, the cascade-based input-doubling classifier increases execution time by over 31 times compared to the Random Forest algorithm, but it remains more than 50 times faster than the baseline input-doubling classifier.

Since the first algorithmic implementation of the cascade-based input-doubling method is more than 30 points more accurate than the baseline input-doubling classifier (according to the Matthews Correlation Coefficient and Cohen's Kappa) and is 50 times faster, we believe it can replace the existing method and be used in practice to solve the problem addressed in this paper.

From a medical expert's perspective, the following conclusions can be drawn based on the obtained results. The application of artificial intelligence methods to solve the given problem is not a new topic. Several studies have focused on addressing this task using well-known machine learning methods [44–47]. However, unlike these works, the authors

have enhanced the classification method specifically for analyzing small datasets, which are quite common in the medical field for various reasons. The three algorithmic implementations of the improved method developed by the authors in this article demonstrate a significant increase in accuracy when solving the task, compared to existing methods. The trade-off for this improvement is the increased duration of the training procedure, which is notably longer than using basic machine learning methods. However, considering the specifics of the task—namely, predicting survival in allogeneic bone marrow transplants—where accuracy is the dominant factor rather than speed, the developed method should be implemented in practice.

Additionally, the software solution developed by the authors operates in an automated mode and does not require high expertise in machine learning from medical staff. This ensures its potential for use by medical personnel with varying levels of training.

Among the potential directions for future research, the following should be considered:

- Investigate the effectiveness of using simpler optimization methods [4] to reduce the execution time of the algorithm while maintaining high accuracy.
- Explore alternative methods for analyzing small datasets to potentially replace ML-1 or ML-2 of the proposed method with other models for the enhanced method implementation described in this paper. For instance, substituting ML-2 with RBF ANN [48] could further improve the method's overall accuracy and decrease its training time.
- Refine the data augmentation procedure by incorporating clustering techniques [31] to significantly reduce the size of the augmented dataset while preserving high accuracy.
- Enhance the proposed method by expanding the input data space using a set of probabilities of class membership [13], rather than class labels as done in this study. This approach may improve the method's accuracy.

6. Conclusions and Future Work

In transplantology, where medical decisions rely heavily on complex data analysis, the challenge of small datasets has become increasingly significant. This field, focused on organ and tissue transplantation, demands exceptional accuracy and precision in predicting outcomes, assessing risks, and tailoring treatment plans. However, the limitations of small datasets pose substantial challenges. This paper introduces a cascade-based input-doubling classifier for predicting survival in allogeneic bone marrow transplants when data are limited.

The method was modeled using a small, imbalanced dataset available in a public repository. The authors investigated four different algorithmic implementations of the proposed method. The accuracy of the classification task was evaluated using several metrics, with particular emphasis on F1-score, Matthews Correlation Coefficient (MCC), and Cohen's Kappa score.

Given that the method utilizes two machine learning algorithms of ANNs, the search for its optimal parameters was performed using the Dual Annealing optimization method. The improved classifier was compared to existing input-doubling classifiers and several other existing methods from different classes. Optimal parameters for all investigated methods were determined using the Dual Annealing optimizer.

The results showed that the first algorithmic implementation of the enhanced input-doubling classifier, which incorporated PNN and the Random Forest algorithm for the training procedure, outperformed all other methods across the investigated metrics. For example, this algorithm is more than 30 points more accurate than the baseline input-doubling classifier (based on the Matthews Correlation Coefficient and Cohen's

Kappa) and is 50 times faster, which underscores its practical value in solving the problem of predicting survival in allogeneic bone marrow transplants when data are limited.

Future research should explore more efficient optimization methods to reduce execution time and investigate alternative data analysis techniques to further enhance its accuracy.

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

Funding: This research was funded by the European Union (through the EURIZON H2020 project, grant agreement 871072).

Data Availability Statement: The dataset used in this study is available here: Ref. [36].

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

Appendix A

Table A1 summarizes the results of the enhanced method for each distance metric studied, as derived from Equations (1)–(6). Note that these results reflect the optimal sigma value for each specific experiment.

Table A1. Results of optimal distance selection for PNN implementation in the proposed method.

Distance	Accuracy	Precision	Recall	F1-Score	Matthews Correlation Coefficient	Cohen's Kappa Score
Euclidean	0.895	0.904	0.91	0.905	0.79	0.79
Manhattan	0.763	0.667	0.875	0.757	0.553	0.534
cosine	0.842	0.952	0.800	0.87	0.69	0.67
Chebyshev	0.789	0.857	0.783	0.818	0.573	0.569
Minkowski	0.842	0.86	0.857	0.858	0.681	0.681
Canberra	0.789	0.905	0.76	0.826	0.578	0.564

Table A2. The meaning of the optimal operating parameters for four algorithmic implementations of the cascade-based input-doubling method.

Algorithm	Parameter's Value
PNN with RandomForest	{'Chebyshev distance', 'sigma=0.508605399', ''n_estimators': 165, 'max_depth': 14}
PNN with XGBoost	{'Euclidian distance', 'sigma=2.69521', 'n_estimators': 179, 'learning_rate': 0.8668009441186528, 'max_depth': 3}
PNN with HistGradientBoosting	{'cosine distance', 'sigma=5.167324462', 'max_iter': 52, 'learning_rate': 0.5598098557162532, 'max_depth': 5}
PNN-1 with PNN-2	{'Euclidian distance', 'sigma1=2.69521'; 'sigma2=4.11678'}

References

1. Tolstyak, Y.; Chopyak, V.; Havryliuk, M. An Investigation of the Primary Immunosuppressive Therapy's Influence on Kidney Transplant Survival at One Month after Transplantation. *Transpl. Immunol.* **2023**, *78*, 101832. [CrossRef]
2. Tolstyak, Y.; Zhuk, R.; Yakovlev, I.; Shakhovska, N.; Gregus Ml, M.; Chopyak, V.; Melnykova, N. The Ensembles of Machine Learning Methods for Survival Predicting after Kidney Transplantation. *Appl. Sci.* **2021**, *11*, 10380. [CrossRef]
3. Bhat, M.; Rabindranath, M.; Chara, B.S.; Simonetto, D.A. Artificial Intelligence, Machine Learning, and Deep Learning in Liver Transplantation. *J. Hepatol.* **2023**, *78*, 1216–1233. [CrossRef] [PubMed]
4. Havryliuk, M.; Hovdysh, N.; Tolstyak, Y.; Chopyak, V.; Kustra, N. Investigation of PNN Optimization Methods to Improve Classification Performance in Transplantation Medicine. In Proceedings of the IDDM'2023: 6th International Conference on Informatics & Data-Driven Medicine, CEUR-WS.org 3609, Bratislava, Slovakia, 17–19 November 2023; pp. 338–345.

5. Huang, S.; Deng, H. *Data Analytics: A Small Data Approach*, 1st ed.; Chapman & Hall/CRC Data Science Series; CRC Press: Boca Raton, FL, USA, 2021; ISBN 978-0-367-60950-4.
6. Krak, I.; Kuznetsov, V.; Kondratiuk, S.; Azarova, L.; Barmak, O.; Padiuk, P. Analysis of Deep Learning Methods in Adaptation to the Small Data Problem Solving. In *Lecture Notes in Data Engineering, Computational Intelligence, and Decision Making*; Babichev, S., Lytvynenko, V., Eds.; Lecture Notes on Data Engineering and Communications Technologies; Springer International Publishing: Cham, Switzerland, 2023; Volume 149, pp. 333–352. ISBN 978-3-031-16202-2.
7. Medykovskvi, M.; Tsmots, I.; Skorokhoda, O. Spectrum Neural Network Filtration Technology for Improving the Forecast Accuracy of Dynamic Processes in Economics. *Actual. Probl. Econ.* **2014**, *162*, 410–416.
8. Izonin, I.; Tkachenko, R.; Berezsky, O.; Krak, I.; Kováč, M.; Fedorchuk, M. Improvement of the ANN-Based Prediction Technology for Extremely Small Biomedical Data Analysis. *Technologies* **2024**, *12*, 112. [CrossRef]
9. Chumachenko, D.; Butkevych, M.; Lode, D.; Frohme, M.; Schmailzl, K.J.G.; Nechyporenko, A. Machine Learning Methods in Predicting Patients with Suspected Myocardial Infarction Based on Short-Time HRV Data. *Sensors* **2022**, *22*, 7033. [CrossRef]
10. Chumachenko, D.; Piletskiy, P.; Sukhorukova, M.; Chumachenko, T. Predictive Model of Lyme Disease Epidemic Process Using Machine Learning Approach. *Appl. Sci.* **2022**, *12*, 4282. [CrossRef]
11. Shaikhina, T.; Lowe, D.; Daga, S.; Briggs, D.; Higgins, R.; Khovanova, N. Machine Learning for Predictive Modelling Based on Small Data in Biomedical Engineering. *IFAC-Pap.* **2015**, *48*, 469–474. [CrossRef]
12. Specht, D.F. Probabilistic Neural Networks. *Neural. Netw.* **1990**, *3*, 109–118. [CrossRef]
13. Zub, K.; Zhezhnych, P.; Strauss, C. Two-Stage PNN–SVM Ensemble for Higher Education Admission Prediction. *BDCC* **2023**, *7*, 83. [CrossRef]
14. Izonin, I.; Tkachenko, R.; Ryvak, L.; Zub, K.; Rashkevych, M.; Pavliuk, O. Addressing Medical Diagnostics Issues: Essential Aspects of the PNN-Based Approach. In Proceedings of the CEUR-WS.org, Volume 2753: Proceedings of the 3rd International Conference on Informatics & Data-Driven Medicine, Växjö, Sweden, 19–21 November 2020; pp. 209–218.
15. Izonin, I.; Tkachenko, R.; Havryliuk, M.; Gregus, M.; Yendyk, P.; Tolstyak, Y. An Adaptation of the Input Doubling Method for Solving Classification Tasks in Case of Small Data Processing. *Procedia Comput. Sci.* **2024**, *241*, 171–178. [CrossRef]
16. Snow, D. *DeltaPy: A Framework for Tabular Data Augmentation in Python*; Social Science Research Network: Rochester, NY, USA, 2020.
17. Xu, L.; Skoularidou, M.; Cuesta-Infante, A.; Veeramachaneni, K. Modeling Tabular Data Using Conditional GAN. *arXiv* **2019**, arXiv:1907.00503.
18. Deep Learning for Tabular Data Augmentation. Available online: https://lschmiddey.github.io/fastpages_/2021/04/10/DeepLearning_TabularDataAugmentation.html (accessed on 16 May 2021).
19. Izonin, I.; Tkachenko, R.; Pidkostelnyi, R.; Pavliuk, O.; Khavalko, V.; Batyuk, A. Experimental Evaluation of the Effectiveness of ANN-Based Numerical Data Augmentation Methods for Diagnostics Tasks. In Proceedings of the 4th International Conference on Informatics & Data-Driven Medicine, Valencia, Spain, 19 November 2021; Volume 3038, pp. 223–232.
20. Nanni, L.; Brahnam, S.; Loreggia, A.; Barcellona, L. Heterogeneous Ensemble for Medical Data Classification. *Analytics* **2023**, *2*, 676–693. [CrossRef]
21. Subbotin, S. Radial-Basis Function Neural Network Synthesis on the Basis of Decision Tree. *Opt. Mem. Neural Netw.* **2020**, *29*, 7–18. [CrossRef]
22. Rokach, L. Taxonomy for Characterizing Ensemble Methods in Classification Tasks: A Review and Annotated Bibliography. *Comput. Stat. Data Anal.* **2009**, *53*, 4046–4072. [CrossRef]
23. Yaman, M.A.; Rattay, F.; Subasi, A. Comparison of Bagging and Boosting Ensemble Machine Learning Methods for Face Recognition. *Procedia Comput. Sci.* **2021**, *194*, 202–209. [CrossRef]
24. Lee, S.-J.; Tseng, C.-H.; Yang, H.-Y.; Jin, X.; Jiang, Q.; Pu, B.; Hu, W.-H.; Liu, D.-R.; Huang, Y.; Zhao, N. Random RotBoost: An Ensemble Classification Method Based on Rotation Forest and AdaBoost in Random Subsets and Its Application to Clinical Decision Support. *Entropy* **2022**, *24*, 617. [CrossRef]
25. Bagnall, A.; Flynn, M.; Large, J.; Line, J.; Bostrom, A.; Cawley, G. Is Rotation Forest the Best Classifier for Problems with Continuous Features? *arXiv* **2018**, arXiv:1809.06705.
26. Bodyanskiy, Y.; Zaychenko, Y.; Pliss, I.; Chala, O. Matrix Neural Network with Kernel Activation Function and Its Online Combined Learning. In Proceedings of the 2022 IEEE 3rd International Conference on System Analysis & Intelligent Computing (SAIC), Kyiv, Ukraine, 4 October 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–4.
27. Kandel, I.; Castelli, M.; Popovič, A. Comparing Stacking Ensemble Techniques to Improve Musculoskeletal Fracture Image Classification. *J. Imaging* **2021**, *7*, 100. [CrossRef]
28. Swaroop, K.; Cheruku, R.; Edla, D.R. Cascading of RBFN, PNN and SVM for Improved Type-2 Diabetes Prediction Accuracy. *Aust. J. Wirel. Technol. Mobil. Secur.* **2019**, *1*, 4.
29. Paul, S. Ensemble Learning—Bagging, Boosting, Stacking and Cascading Classifiers in Machine Learning. Available online: https://medium.com/@saugata.paul1010/ensemble-learning-bagging-boosting-stacking-and-cascading-classifiers-in-machine-learning-9c66cb271674 (accessed on 2 October 2022).

30. Fernández-Alemán, J.L.; Carrillo-de-Gea, J.M.; Hosni, M.; Idri, A.; García-Mateos, G. Homogeneous and Heterogeneous Ensemble Classification Methods in Diabetes Disease: A Review. In Proceedings of the 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 23–27 July 2019; pp. 3956–3959.
31. Izonin, I.; Tkachenko, R.; Yemets, K.; Gregus, M.; Tomashy, Y.; Pliss, I. An Approach Towards Reducing Training Time of the Input Doubling Method via Clustering for Middle-Sized Data Analysis. *Procedia Comput. Sci.* **2024**, *241*, 32–39. [CrossRef]
32. Bodyanskiy, Y.V.; Tyshchenko, O.K. A Hybrid Cascade Neural Network with Ensembles of Extended Neo-Fuzzy Neurons and Its Deep Learning. In Proceedings of the Information Technology, Systems Research, and Computational Physics, Cracow, Poland, 2–5 July 2018; Springer: Cham, Switzerland, 2018; pp. 164–174.
33. García-Pedrajas, N.; Ortiz-Boyer, D.; del Castillo-Gomariz, R.; Hervás-Martínez, C. Cascade Ensembles. In Proceedings of the Computational Intelligence and Bioinspired Systems, Barcelona, Spain, 8–10 June 2005; Springer: Berlin/Heidelberg, Germany, 2005; pp. 598–603.
34. Izonin, I.; Kazantzi, A.K.; Tkachenko, R.; Mitoulis, S.-A. GRNN-Based Cascade Ensemble Model for Non-Destructive Damage State Identification: Small Data Approach. *Eng. J.* 2024; *under review*.
35. Samuelson, F.; Brown, D.G. Application of Cover's Theorem to the Evaluation of the Performance of CI Observers. In Proceedings of the The 2011 International Joint Conference on Neural Networks, San Jose, CA, USA, 31 July–5 August 2011; IEEE: San Jose, CA, USA, 2011; pp. 1020–1026.
36. Bone Marrow Transplant: Children—UCI Machine Learning Repository. Available online: https://archive.ics.uci.edu/dataset/565/bone+marrow+transplant+children (accessed on 25 November 2023).
37. Gudyś, A.; Sikora, M.; Wróbel, Ł. RuleKit: A Comprehensive Suite for Rule-Based Learning. *Knowl. Based Syst.* **2020**, *194*, 105480. [CrossRef]
38. Sikora, M.; Wróbel, Ł.; Gudyś, A. GuideR: A Guided Separate-and-Conquer Rule Learning in Classification, Regression, and Survival Settings. *Knowl. Based Syst.* **2019**, *173*, 1–14. [CrossRef]
39. Wróbel, Ł.; Gudyś, A.; Sikora, M. Learning Rule Sets from Survival Data. *BMC Bioinform.* **2017**, *18*, 285. [CrossRef]
40. Kałwak, K.; Porwolik, J.; Mielcarek, M.; Gorczyńska, E.; Owoc-Lempach, J.; Ussowicz, M.; Dyla, A.; Musiał, J.; Paździor, D.; Turkiewicz, D.; et al. Higher CD34+ and CD3+ Cell Doses in the Graft Promote Long-Term Survival, and Have No Impact on the Incidence of Severe Acute or Chronic Graft-versus-Host Disease after In Vivo T Cell-Depleted Unrelated Donor Hematopoietic Stem Cell Transplantation in Children. *Biol. Blood Marrow Transplant.* **2010**, *16*, 1388–1401. [CrossRef] [PubMed]
41. Manna, S. Small Sample Estimation of Classification Metrics. In Proceedings of the 2022 Interdisciplinary Research in Technology and Management (IRTM), Kolkata, India, 24–26 February 2022; IEEE: Kolkata, India, 2022; pp. 1–3.
42. Berezsky, O.; Pitsun, O.; Liashchynskyi, P.; Derysh, B.; Batryn, N. Computational Intelligence in Medicine. In *Lecture Notes in Data Engineering, Computational Intelligence, and Decision Making*; Babichev, S., Lytvynenko, V., Eds.; Lecture Notes on Data Engineering and Communications Technologies; Springer International Publishing: Cham, Switzerland, 2023; Volume 149, pp. 488–510. ISBN 978-3-031-16202-2.
43. Ryczkowski, A.; Piotrowski, T.; Staszczak, M.; Wiktorowicz, M.; Adrich, P. Optimization of the Regularization Parameter in the Dual Annealing Method Used for the Reconstruction of Energy Spectrum of Electron Beam Generated by the AQURE Mobile Accelerator. *Z. Für Med. Phys.* **2023**, *34*, 510–520. [CrossRef]
44. Chadaga, K.; Prabhu, S.; Sampathila, N.; Chadaga, R. A Machine Learning and Explainable Artificial Intelligence Approach for Predicting the Efficacy of Hematopoietic Stem Cell Transplant in Pediatric Patients. *Healthc. Anal.* **2023**, *3*, 100170. [CrossRef]
45. Gross, M.-P.; Taormina, R.; Cominola, A. A Machine Learning-Based Framework and Open-Source Software for Non Intrusive Water Monitoring. *Environ. Model. Softw.* **2025**, *183*, 106247. [CrossRef]
46. Yu, T.-C.; Yang, C.-K.; Hsu, W.-H.; Hsu, C.-A.; Wang, H.-C.; Hsiao, H.-J.; Chao, H.-L.; Hsieh, H.-P.; Wu, J.-R.; Tsai, Y.-C.; et al. A Machine-Learning-Based Algorithm for Bone Marrow Cell Differential Counting. *Int. J. Med. Inform.* **2025**, *194*, 105692. [CrossRef]
47. Buturovic, L.; Shelton, J.; Spellman, S.R.; Wang, T.; Friedman, L.; Loftus, D.; Hesterberg, L.; Woodring, T.; Fleischhauer, K.; Hsu, K.C.; et al. Evaluation of a Machine Learning-Based Prognostic Model for Unrelated Hematopoietic Cell Transplantation Donor Selection. *Biol. Blood Marrow Transplant.* **2018**, *24*, 1299–1306. [CrossRef]
48. Shaikhina, T.; Khovanova, N.A. Handling Limited Datasets with Neural Networks in Medical Applications: A Small-Data Approach. *Artif. Intell. Med.* **2017**, *75*, 51–63. [CrossRef] [PubMed]