electronics

MDPI

Article

Transferring Black-Box Decision Making to a White-Box Model

Bojan Žlahtič [1,*], Jernej Završnik [2,3,4,5], Helena Blažun Vošner [2,3,6] and Peter Kokol [1,2]

[1] Faculty of Electrical Engineering and Computer Science, University of Maribor, 2000 Maribor, Slovenia
[2] Community Healthcare Center Dr. Adolf Drolc Maribor, 2000 Maribor, Slovenia
[3] Alma Mater Europaea—ECM, 2000 Maribor, Slovenia
[4] Science and Research Center Koper, 6000 Koper, Slovenia
[5] Faculty of Natural Sciences and Mathematics, University of Maribor, 2000 Maribor, Slovenia
[6] Faculty of Health and Social Sciences, 2380 Slovenj Gradec, Slovenia
* Correspondence: bojan.zlahtic@um.si

Abstract: In the rapidly evolving realm of artificial intelligence (AI), black-box algorithms have exhibited outstanding performance. However, their opaque nature poses challenges in fields like medicine, where the clarity of the decision-making processes is crucial for ensuring trust. Addressing this need, the study aimed to augment these algorithms with explainable AI (XAI) features to enhance transparency. A novel approach was employed, contrasting the decision-making patterns of black-box and white-box models. Where discrepancies were noted, training data were refined to align a white-box model's decisions closer to its black-box counterpart. Testing this methodology on three distinct medical datasets revealed consistent correlations between the adapted white-box models and their black-box analogs. Notably, integrating this strategy with established methods like local interpretable model-agnostic explanations (LIMEs) and SHapley Additive exPlanations (SHAPs) further enhanced transparency, underscoring the potential value of decision trees as a favored white-box algorithm in medicine due to its inherent explanatory capabilities. The findings highlight a promising path for the integration of the performance of black-box algorithms with the necessity for transparency in critical decision-making domains.

Keywords: explainable artificial intelligence; XAI; machine learning; deep learning; white box; black box

check for updates

Citation: Žlahtič, B.; Završnik, J.; Blažun Vošner, H.; Kokol, P. Transferring Black-Box Decision Making to a White-Box Model. *Electronics* **2024**, *13*, 1895. https://doi.org/10.3390/electronics13101895

Academic Editor: Gabriella Olmo

Received: 5 April 2024
Revised: 1 May 2024
Accepted: 10 May 2024
Published: 12 May 2024

1. Introduction

Medicine and healthcare experts continually incorporate new artificial intelligence and machine learning approaches, algorithms, and concepts into all facets of the fundamental building blocks of their organizational structures [1–5] to provide the best possible care for billions of patients worldwide. With the rapidly expanding knowledge base in the field of medicine and the enormous amounts of patient data collected annually, there are not enough professional human resources available to manage all the data, let alone make proper use of it [6]. Machine learning algorithms can serve as invaluable decision support systems and knowledge extraction tools for medical professionals. By leveraging these technologies, healthcare providers can efficiently manage the vast quantities of data they accumulate and attend to the continuous stream of patients with professionalism and equitability [7]. Machine learning in medicine is being used in genetics [8], medical imaging [9], radiology [10], cancer prediction [11], 2D echocardiology [12], and neurosurgical outcome prediction [13], and the list goes on because everywhere where there are data created, there is a potential to apply machine learning.

The black-box nature of a machine learning algorithm makes it extremely difficult for their end users to interpret the results [14]. A black-box [15] algorithm is a computational model or system for which its internal workings and processes are not readily accessible or understandable to an observer. The underlying mechanisms or decision-making processes are hidden, but the input–output relationship is known. A white-box [16] algorithm is a computational model or system that provides transparency and interpretability in its internal workings and decision-making processes. Unlike black-box algorithms, white-box algorithms are designed to be easily understandable and interpretable by humans. They often have a clear and explicit representation of the logic and rules used to arrive at their outputs. In certain areas of application, this creates a chasm between top-tier classification algorithms and users demanding transparency since the algorithm's results may be applied in a life-or-death scenario [17]. Modern black-box deep learning algorithms reached a point where they can accomplish complex decision-making tasks with unprecedented accuracy on an enormous scale. This, however, presents a great dilemma: can those decisions be trusted? On the other hand, the few white-box classifiers in use cannot measure up to those black-box classifiers [18]. Medicine and autonomous vehicles [19,20] are two fields of application where the ability to explain decisions presents a high-risk scenario. Recently, a new research field has been gaining traction, explainable artificial intelligence (XAI) [21–25], where the aim is to make black-box machine learning algorithms trustable in terms of transparency and the ability to interpret their results. The thus far proposed solutions, which are few and far between, are focused on adding a transparency layer via attribute analysis in combination with the black-box classifier's decision, which is the case with SHapley Additive exPlanations (SHAPs) [26] or local interpretable model-agnostic explanation (LIMEs) [27]. Those present two of the most prominent approaches in medicine, but since no algorithm to date has been successful in presenting a realistic and complete solution to the problems that were outlined by XAI, any novel approach that can provide a glance into black-box models should be taken into consideration.

The foundational objective of XAI, as defined by the Defense Advanced Research Projects Agency (DARPA) [28], is to bridge the divide between learning performance and explainability in AI systems. Two polarities are presented in this context: neural networks, representing the peak of learning performance, and decision trees, representing the pinnacle of explainability. The space between these two paradigms embodies the core challenge that XAI attempts to overcome, as seen in Figure 1 where the strengths and weaknesses of both black-box and white-box algorithms clearly outline the gray area in the field of artificial intelligence. This involves either harnessing enhanced performance to facilitate greater accuracy or utilizing heightened explainability to increase usability and trustability. The prioritization of these aspects varies across different application domains. Notably, in the field of medicine, for an array of reasons [29], explainability is currently of greater significance than plain performance.

In collaborative efforts with medical professionals in the data science field, a pronounced demand was identified for elucidating the underlying rationale of decision-making processes. Many experts were hesitant to adopt solutions in their domains that lacked clear interpretative capabilities. In the course of developing a novel white-box machine learning algorithm, termed "Data Canyons" [30], it was discerned that specific data instances obfuscated the clarity of the decision-making rationale. This obfuscation was identifiable when contrasting the classification outcomes of different machine learning methodologies. Consequently, an approach emerged: the isolation and relabeling of these ambiguous data entries. This gave rise to the concept of knowledge transfer between a black-box model, which possessed the capability to neglect certain instances, and a transparent white-box model.

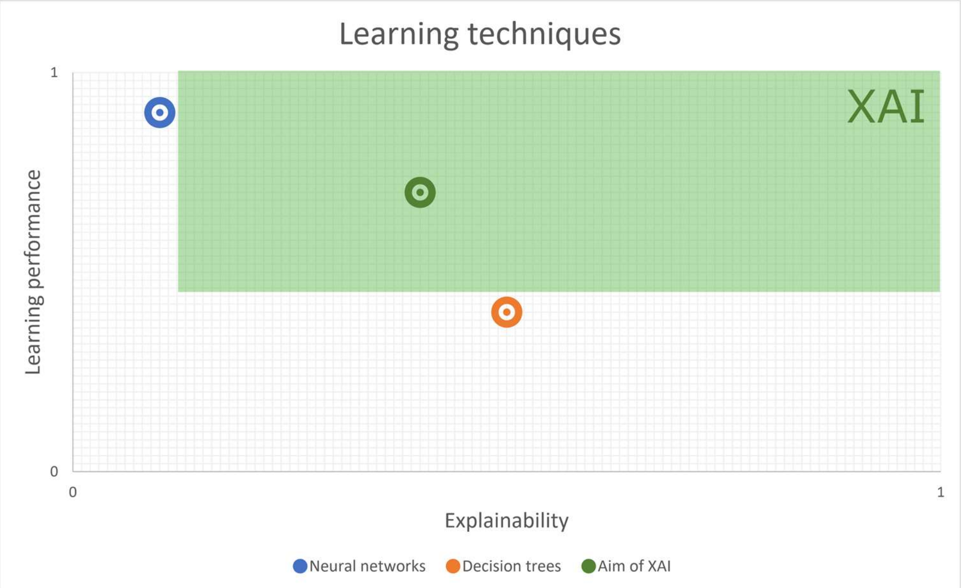


Figure 1. The two opposite polarities outlined by DARPA and the gap between them that is the focus of XAI.

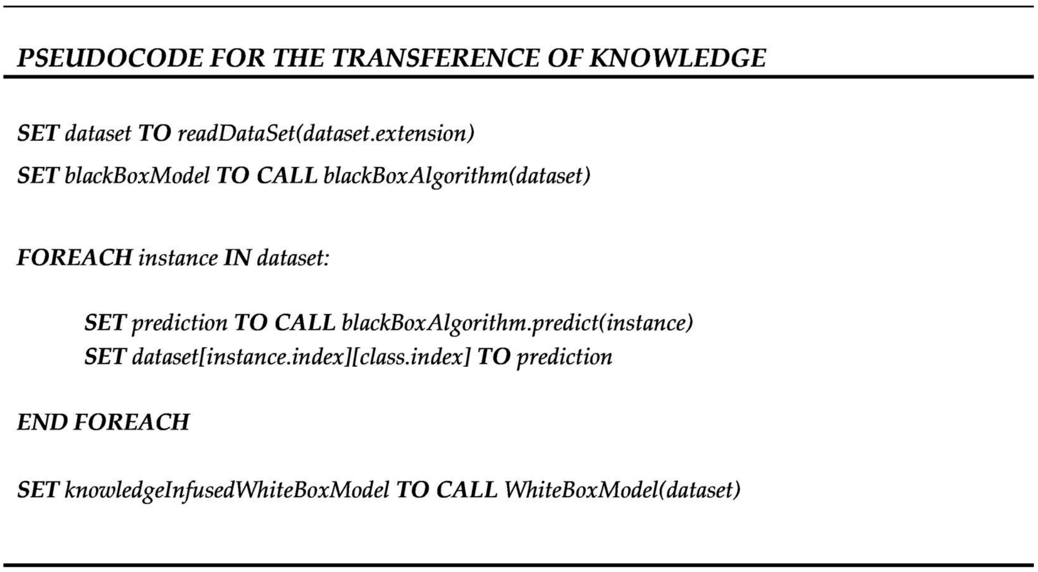
By stacking new methods to translate the black-box models or at least to interpret them, it might be possible to "open" the black-box [31]. The primary emphasis of prevailing XAI techniques lies in presenting the attributes that play a significant role in the decision-making mechanism of a black-box algorithm [32]. However, the proposed methodology goes beyond this by elucidating the rationale behind the decisions made. This study introduces a pioneering methodology, believed to be the first of its kind, designed specifically for application in the fields of medicine and healthcare. This approach can be synergistically utilized alongside established methods such as LIME and SHAP to enhance transparency and foster trust in the decision-making processes of black-box algorithms. In the domain of medicine, the reasoning process holds greater importance compared to the mere identification of involved attributes [33]. Specifically, decision trees [34] garner considerable attention in the medical and healthcare fields [35] due to their widespread adoption as a machine learning approach, primarily owing to their inherent capacity for explanatory visual representation. With the presented method, an effort has been made to transfer part of the decision-making process of a black-box algorithm to a white-box decision-tree algorithm by incorporating the differences in the initial decision-making process of both algorithms into the learning phase of the white-box algorithm.

2. Materials and Methods

Using a machine learning test bench like Weka [36,37] allows for testing many algorithms at the same time and finding the algorithms that give the most promising results. Using this technique, it was observed that in certain data types or instance distributions relating to the instance classes, there is a significant difference between how accurate those algorithms perform and how accurately they perform for specific classes in the dataset. This difference is, of course, vital because of the need to use the most stable and accurate classification algorithm, but if the best results are given by an algorithm that is of the black-box variety, this raises concerns with respect to trustability and interpretability. If there is a vast difference between the preferred performance of the black-box model compared to the results of the white-box model and the reason for this is the distribution of instances relating to the class, then there might be a way to increase the performance of the white-box algorithm by utilizing the black-box algorithms model, and with that, open a window into the black box. This method is not applicable in every scenario; however, it could be a staple in the toolbox of data science and machine learning. The same principle could be used utilizing the powers of any ensemble method or any voting-driven prediction construct since the base idea utilizes the final result of the predictions of any approach. The black-box classification results are introduced into the original data and used as the basis for the white-box classifier learning process. If the distribution of data is right and there is a need to correctly classify classes that are in the minority, this method could create a white-box algorithm that includes the reasoning behind the decisions of the black-box algorithm. The base principle, therefore, is the infusion of knowledge using the results of a black-box algorithm, giving desirable outputs to a white-box algorithm that, in using the results of the black-box algorithm, creates a better-performing model in the sense of the desired outcome. The white-box classifier can then be interpreted using its visualization method. The scenario used throughout this paper is a medical dataset where there is a vast discrepancy between the distribution of classes presented in the train data. The assumption is that most classification approaches will yield a satisfactory outcome but at the cost of the misclassification of underrepresented classes. However, some algorithms will perform better than others when it comes to those underrepresented classes and will, therefore, present the entry point into the transference of learning between black-box and white-box approaches. In medicine, these data imbalances are common, and more often than not, the critical part of the classification emphasis lies on the underrepresented classes.

Weka, originating from the University of Waikato in New Zealand, is an open-source software suite that offers a comprehensive collection of machine learning algorithms for data mining tasks, encompassing classification, regression, clustering, association rules, and data preprocessing. Implemented in Java [38,39], a high-level, object-oriented programming language, this platform is employed in both educational and research contexts, providing tools for data visualization, analysis, and predictive modeling. All selected datasets were evaluated using Weka and all algorithms that were able to work with the presented data. Moreover, for the sake of repeatability, the default settings were used for each algorithm. The version of Weka used was 3.8.6, and the version of Java was 17.0.6. The data that were used for testing were the data provided via the Weka installation, and they are medical in nature. The used datasets were “breast-cancer.arff”, “diabetes.arff”, and “hypothyroid.arff”. The breast cancer domain was obtained from the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia. We thank M. Zwitter and M. Soklic for providing the data. The diabetes dataset is the Pima Indians diabetes dataset from the National Institute of Diabetes and Digestive and Kidney Diseases. The thyroid disease records were supplied by the Garavan Institute and J. Ross Quinlan, New South Wales Institute, Sydney, Australia.

All datasets were tested with all applicable algorithms, and all results were obtained in order to find a deviation on which the presented theory could be applied. In all cases, the best-performing machine learning model that detected the most minority instances was selected to transfer its knowledge to the best-performing decision tree algorithm. In the second part, the results of the algorithms without default visualization have been used to create a new model using decision trees to try to transfer the decision process of one algorithm to another. The base process is shown in Figure 2. To begin, the dataset was loaded, and a black-box algorithm was utilized to construct an initial model. This model generated predictions for each instance within the dataset. Subsequently, these predictions were used to update the dataset by incorporating the predicted outcomes into their respective instances. Following this, a white-box model was created by applying a white-box algorithm to the newly modified dataset. The resulting model presents the final knowledge-infused white-box model. In the final stage, the original output of the decision tree algorithms was compared to the new output on the original dataset to evaluate the change in performance.



```
PSEUDOCODE FOR THE TRANSFERENCE OF KNOWLEDGE

SET dataset TO readDataSet(dataset.extension)
SET blackBoxModel TO CALL blackBoxAlgorithm(dataset)

FOREACH instance IN dataset:

    SET prediction TO CALL blackBoxAlgorithm.predict(instance)
    SET dataset[instance.index][class.index] TO prediction

END FOREACH

SET knowledgeInfusedWhiteBoxModel TO CALL WhiteBoxModel(dataset)
```

Figure 2. Pseudocode for the process of knowledge transference and the creation of the knowledge-infused white-box model.

3. Results

Comprehensive testing was conducted on all three datasets using the available classifiers in Weka. However, certain classifiers were excluded from the analysis due to incompatible datatypes within the datasets or dependencies of the classifiers. The primary objective of testing multiple classifiers was to identify a black-box classifier that performed favorably and a white-box classifier that exhibited suboptimal performance concerning the desired output, specifically the accurate classification of the minority class in unbalanced datasets. In the context of unbalanced datasets, where one class is significantly more prevalent than the others (e.g., one class appears significantly more frequently than the other), classifiers tend to prioritize the more frequent class. However, in medical applications, it is crucial to focus on the classes that occur less frequently due to the distribution of positive and negative examples in medical datasets. Black-box algorithms often outperform white-box algorithms in such scenarios, necessitating knowledge transfer from black-box models to white-box models. The selection of appropriate black-box and white-box classifiers is demonstrated through the presentation of complete testing results using the breast cancer dataset. As significant performance discrepancies between black-box and white-box classifiers were not exhibited by the other two datasets, the complete test results for all classifiers have been omitted. Instead, focus has been placed on the selected classifiers for each dataset to facilitate comparison and analysis.

The selection of the algorithm for knowledge transfer is contingent upon the desired decision outcome and the appropriate decision outcomes of utilized black-box algorithms. In situations where the dataset is imbalanced, the emphasis lies on accurately classifying the minority classes. In the specific case of the breast cancer database, BayesNet was chosen as it exhibited the highest proficiency in classifying "recurrence-events" while maintaining overall accuracy. Similarly, for the diabetes dataset, NaiveBayes was selected due to its superior performance in predicting the "tested_positive" class. In the case of the thyroid dataset, NaiveBayes was again chosen because it demonstrated the effective detection of the "secondary_hypothyroid" class.

3.1. Breast Cancer Dataset Results

There were 44 algorithms that were able to complete the classification process, as can be seen in Table 1. In all cases, 10-fold cross-validation was used. However, for the purpose of this paper, at the first testing stage, the method of division into the training and testing sets was less critical as long as it stayed consistent throughout the whole first phase with every dataset used. As can be seen from the test results in Table 1, looking only at the correctly classified instances, there is no significant deviation. On the other hand, looking at the decisions in the confusion matrices, it can be seen that the correctly classified instances of the recurrence events range from 0 to 37 out of the 85 that are present in the dataset. Almost all algorithms struggle to identify the class represented in a smaller number. Nine of these algorithms even disregard them altogether and classify all instances as no-recurrence-events. If the emphasis was placed on identifying more of the 85 recurrence events, the nine algorithms would not be usable, at least not without further adjustments. What can also be seen clearly from the results is that correctly classified instances do not present an optimal measure of accuracy, especially if the focus is not the overall classification but rather the detection of data "minorities". To test this theory, an algorithm had to be identified for which its confusion matrix captured the most results from the minority instance group. While none of the algorithms produced particularly satisfactory results, BayesNet was found to correctly classify 37 instances, whereas J48, which includes a visualization layer, managed to classify only 23 of the 85 instances of interest. Consequently, these two algorithms were selected to verify this hypothesis regarding knowledge transference between algorithms. Both models created during the test were preserved. Subsequently, to evaluate the change in accuracy, both models were reloaded, and their performance was retested using the entire dataset as the test set. The results of these tests can be found in Table 2. From the results, it is clearly seen that the focus of the models of those algorithms is on different classes, while, interestingly, they maintain the same percentage of correctly classified instances. To compare the change in the decision tree model, the output decision tree was captured in Figure 3.

Table 1. Initial classification results for the breast cancer dataset.

Nr.	Classifier	Correctly Classified Instances	Nr.	Classifier	Correctly Classified Instances
1.	Naïve Bayes	72.02%	23.	Multiclass Classifier	68.88%
2.	Bayes Net	71.67%	24.	Multiclass Classifier Updateable	69.93%
3.	Naïve Bayes Multinomial Text	70.27%	25.	Multischeme	70.27%
4.	Naïve Bayes Updateable	71.67%	26.	Random Committee	67.48%
5.	Logistic	68.88%	27.	Randomizable Filtered Classifier	66.78%
6.	Multilayer Perceptron	64.68%	28.	Random Subspace	70.27%
7.	SGD	69.93%	29.	Stacking	70.27%
8.	SGD Text	70.27%	30.	Vote	70.27%
9.	Simple Logistic	75.17%	31.	Weighted Instances Handler Wrapper	70.27%
10.	SMO	69.58%	32.	Input Mapped Classifier	70.27%
11.	Voted Perceptron	71.32%	33.	Decision Table	73.42%
12.	IBk	72.37%	34.	JRip	70.97%
13.	KStar	73.42%	35.	OneR	65.73%
14.	LWL	72.37%	36.	PART	71.32%
15.	Ada Boost M1	70.27%	37.	ZeroR	70.27%
16.	Attribute Selected Classifier	73.07%	38.	Decision Stump	68.53%
17.	Bagging	69.23%	39.	Hoeffding Tree	69.93%
18.	Classification Via Regression	71.32%	40.	J48	75.52%
19.	CV Parameter Selection	70.27%	41.	LMT	75.17%
20.	Filtered Classifier	75.52%	42.	Random Forest	69.58%
21.	Iterative Classifier Optimizer	74.82%	43.	Random Tree	66.78%
22.	Logit Boost	72.37%	44.	REP Tree	70.62%

Table 2. Results for the breast cancer dataset.

	Classification Results		
Execution Phases	**Classifier**	**Correctly Classified Instances**	**Confusion Matrix** **a = no-recurrence-events** **b = recurrence-events**
Initial phase (accessing accuracy)	BayesNet (10-fold cross-validation)	71.67%	a b 169 32 a 48 37 b
	J48 (10-fold cross-validation)	75.52%	a b 193 8 a 62 23 b
Second phase (acquiring models)	BayesNet (entire dataset)	75.87%	a b 173 28 a 41 44 b
	J48 (entire dataset)	75.87%	a b 194 7 a 62 23 b
Final phase (accessing new accuracy)	J48 with BayesNet classes (entire dataset)	72.72%	a b 169 32 a 46 39 b

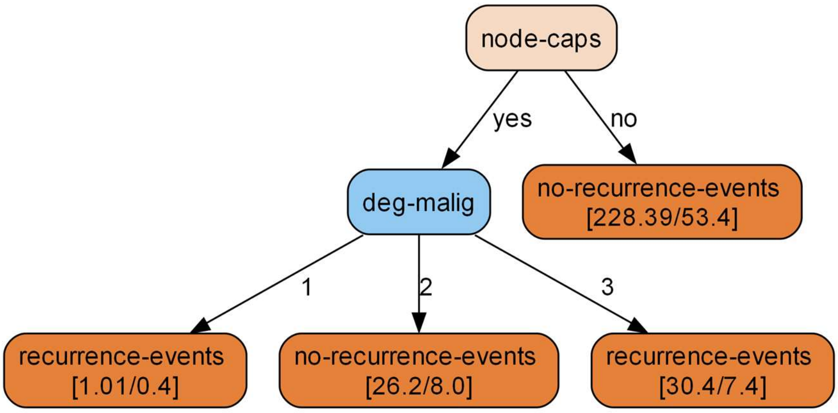

node-caps
yes
no
deg-malig
no-recurrence-events [228.39/53.4]
1
2
3
recurrence-events [1.01/0.4]
no-recurrence-events [26.2/8.0]
recurrence-events [30.4/7.4]

Figure 3. The base decision tree for the breast cancer dataset.

Now that the two models were available, the transference of knowledge was to be applied, necessitating the creation of a new column to replace the old class column of the breast cancer dataset. The knowledge from the BayesNet model was chosen for transfer, and the Weka API was utilized within a Java program to gather all decisions made by the BayesNet model based on the original data. This process yielded all the mappings of the decision making from the actual instance class value to the BayesNet model prediction. Subsequently, the original class values were replaced with the newly gathered values, and the result was saved as a new dataset. This new dataset was then used to create a new model using Weka and the J48 algorithm. It was anticipated that the new model would perform similarly to the BayesNet model; thus, both models were used on the original dataset to assess any changes in performance. The results can be seen in Table 2 under the final phase. It can clearly be seen that the new J48 model behaves differently and that its confusion matrix is more closely related to the model of BayesNet than to the confusion matrix of the base J48 model. Additionally, the decision tree output was compared to the output of the new J48 model, and it was naturally found that the decision tree has changed substantially, as seen in Figure 4. In the last testing step, the correlation of the output of the new J48 model to both the results of BayesNet and the base J48 models was checked. The correlation between the BayesNet results and the base J48 model was 0.563884331, while the correlation between the BayesNet model and the knowledge-infused model J48 was 0.841531005, so the new model has a much higher correlation to the BayesNet model.

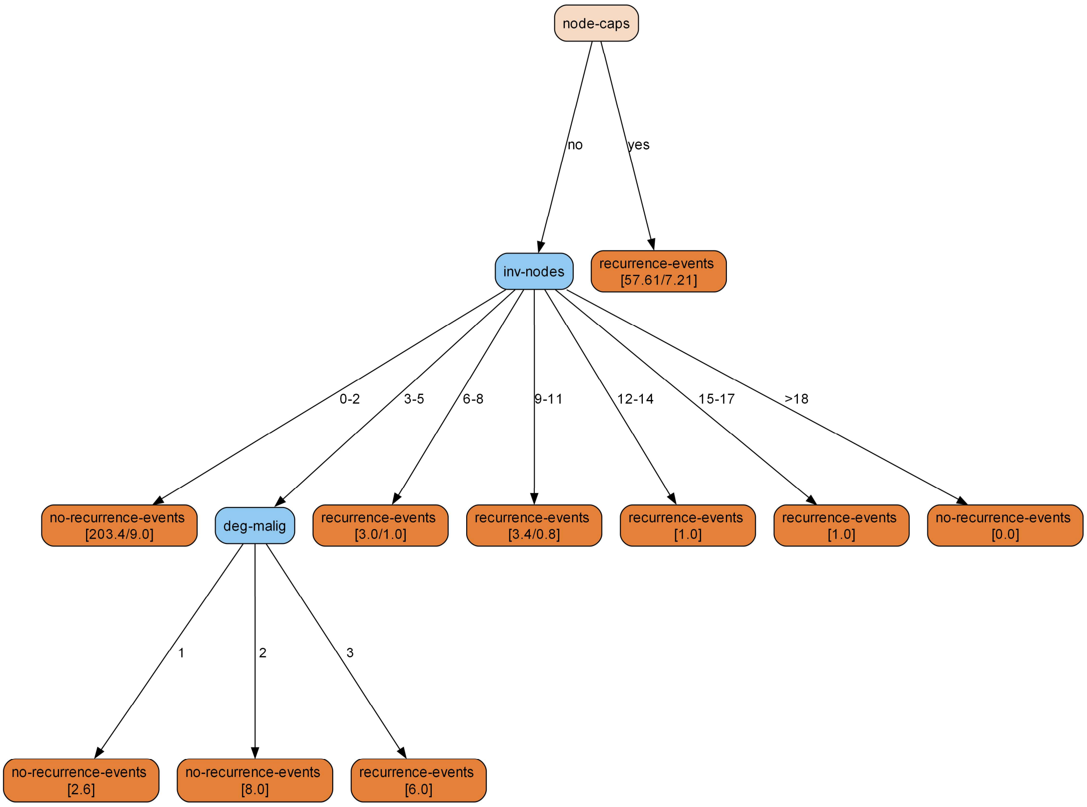

node-caps
no
yes
inv-nodes
recurrence-events [57.61/7.21]
0-2
3-5
6-8
9-11
12-14
15-17
>18
no-recurrence-events [203.4/9.0]
deg-malig
recurrence-events [3.0/1.0]
recurrence-events [3.4/0.8]
recurrence-events [1.0]
recurrence-events [1.0]
no-recurrence-events [0.0]
1
2
3
no-recurrence-events [2.6]
no-recurrence-events [8.0]
recurrence-events [6.0]

Figure 4. The decision tree for the breast cancer dataset with the BayesNet class model.

3.2. Diabetes Dataset Results

To validate the results, the diabetes dataset was employed, and the entire process was repeated. A total of 46 algorithms were run. Notably, nine algorithms misclassified one class and consequently failed to detect any instances of the class “tested_positive”. It is important to highlight that among all three datasets, this one exhibited the smallest imbalance. Bayes classifiers and tree classifiers were selected to compare the results with previous and subsequent testing outcomes. Consequently, in Table 3, under the initial phase, the results of two applicable classifiers are presented. In this scenario, the performance of the J48 classifier was found to be similar to that of the NaiveBayes classifier. However, when both algorithms were further tested against the whole original dataset, a slight shift in performance was observed, as documented in Table 3 under the second phase. At this stage, the outputs of the NaiveBayes model were combined with the whole dataset to create a new J48 model using this dataset. This new model was then tested against the original data, and a change in the results was again noted, as shown in Table 3 under the final phase. With all results accounted for, the correlation was checked, and a strong correlation was again expected between the NaiveBayes results and the NaiveBayes-infused J48 results. A less pronounced correlation was expected between the NaiveBayes and the base J48 model. As anticipated, the correlation between the base NaiveBayes and J48 model was 0.691951407, significantly lower than the correlation between the results of the NaiveBayes model and the knowledge-infused J48 model, which was 0.946443719. This example demonstrated that although the differences in correctly classified instances and the distribution of the confusion matrix were not as marked as in the breast cancer dataset, the decision structure of the J48 model was still significantly reshaped to achieve a strong correlation between the black-box model and the white-box model. The differences in model visualization are clearly evident in Figures 5 and 6.

Table 3. Results for the diabetes dataset.

	Classification Results		
Execution Phases	**Classifier**	**Correctly Classified Instances**	**Confusion Matrix** **a = tested_negative** **b = tested_positive**
Initial phase (accessing accuracy)	NaiveBayes (10-fold cross-validation)	76.30%	$\begin{array}{cc\|c} a & b & \\ \hline 422 & 78 & a \\ 104 & 164 & b \end{array}$
	J48 (10-fold cross-validation)	73.82%	$\begin{array}{cc\|c} a & b & \\ \hline 407 & 93 & a \\ 108 & 160 & b \end{array}$
Second phase (acquiring models)	NaiveBayes (entire dataset)	76.30%	$\begin{array}{cc\|c} a & b & \\ \hline 421 & 79 & a \\ 103 & 165 & b \end{array}$
	J48 (entire dataset)	84.11%	$\begin{array}{cc\|c} a & b & \\ \hline 468 & 32 & a \\ 90 & 178 & b \end{array}$
Final phase (accessing new accuracy)	J48Model with NaiveBayes classes (entire dataset)	76.04%	$\begin{array}{cc\|c} a & b & \\ \hline 417 & 83 & a \\ 101 & 167 & b \end{array}$

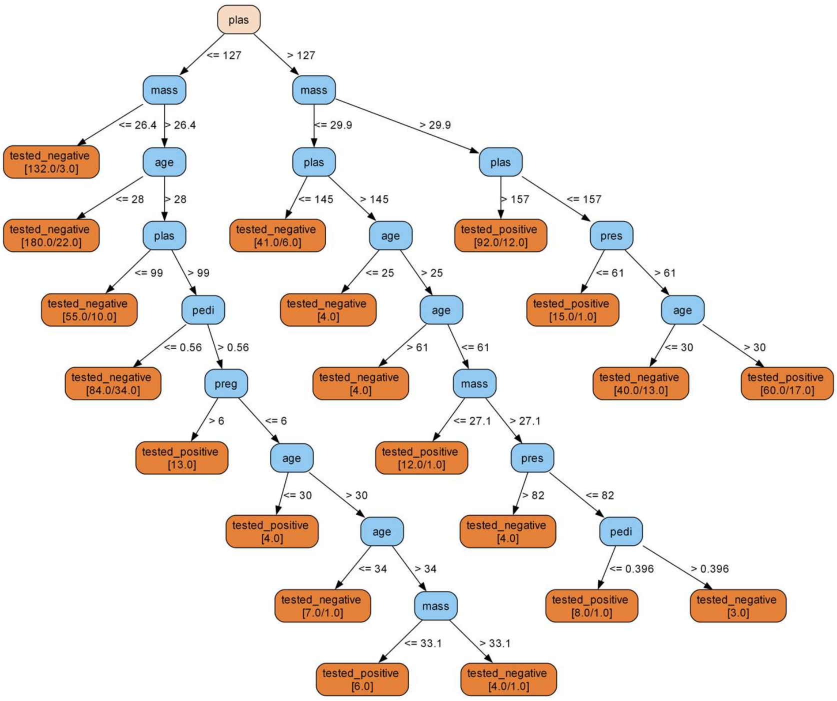


Figure 5. The base decision tree for the diabetes dataset.

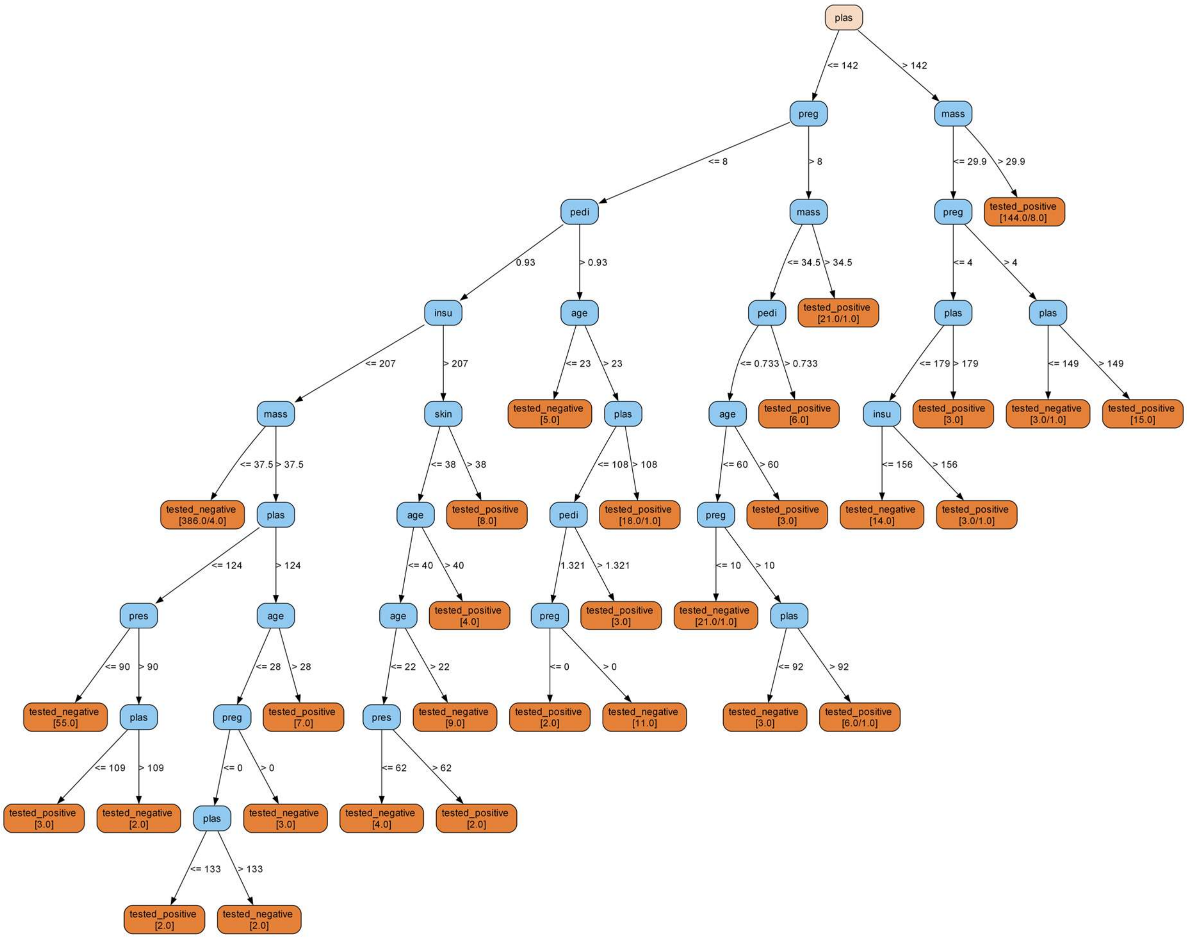


Figure 6. The decision tree for the diabetes dataset with the NaiveBayes classes model.

3.3. *Thyroid Disease Dataset Results*

With the final tested dataset, tests were performed with 41 classifiers from the Weka collection. Similarly to previous findings, some classifiers performed poorly on data present in the minority, as this dataset is also highly unbalanced. Eight classifiers classified all instances as "negative" and misclassified all other instances. This dataset, having the highest imbalance and consisting of four classes unlike the previous datasets, which had only two, presented unique challenges. In the testing process, decision tree classifiers performed very well compared to other classifiers, although the Bayes classifiers were not far behind. Ultimately, the NaiveBayes and J48 classifiers were chosen, with expectations that the transfer of knowledge would not be as significant since the performance differences were not severe, as depicted in Table 4 under the initial phase. In Table 4, under the second phase, it is noted that when the selected models were tested on the original dataset to compare the result of the knowledge-infused model, there was no significant change in the results. Weka API was again used to obtain all classification results for the NaiveBayes classifier, which were then mapped to the original dataset. This newly created dataset was used to develop a new J48 model, the performance of which was tested on the original dataset. As shown in Table 4 under the final phase, the performance of this model changed compared to the original J48 model, as did the decision trees, which are depicted in Figures 7 and 8. It appeared that the performance of the infused model was brought closer to that of the NaiveBayes model, prompting another test of the correlation between the NaiveBayes classification results and the original J48 model results, which yielded a score of 0.73364276, and a value of 0.913280661 was obtained for the correlation between the NaiveBayes model classification results and the knowledge-infused J48 model classification results. These results suggest a much higher correlation between the results of the black-box models and the infused J48 models, indicating that although the differences in the base models did not seem significant, a substantial portion of the decision-making process was successfully transferred from one model to another.

Table 4. Results for the hypothyroid dataset.

Classification Results			
Execution Phases	**Classifier**	**Correctly Classified Instances**	**Confusion Matrix** **a = negative** **b = compensated_hypothyroid** **c = primary_hypothyroid** **d = secondary_hypothyroid**
Initial phase (accessing accuracy)	NaiveBayes (10-fold cross-validation)	95.28%	a b c d 3456 14 9 2 a 130 62 2 0 b 9 10 76 0 c 2 0 0 0 d
	J48 (10-fold cross-validation)	99.57%	a b c d 3476 3 2 0 a 1 191 2 0 b 3 3 89 0 c 2 0 0 0 d
Second phase (acquiring models)	NaiveBayes (entire dataset)	95.44%	a b c d 3454 14 10 3 a 127 65 2 0 b 9 7 79 0 c 0 0 0 2 d
	J48 (entire dataset)	99.81%	a b c d 3480 1 0 0 a 1 193 0 0 b 0 3 92 0 c 2 0 0 0 d
Final phase (accessing new accuracy)	J48 with NaiveBayes classes (entire dataset)	95.28%	a b c d 3458 18 5 0 a 132 58 4 0 b 9 8 78 0 c 2 0 0 0 d

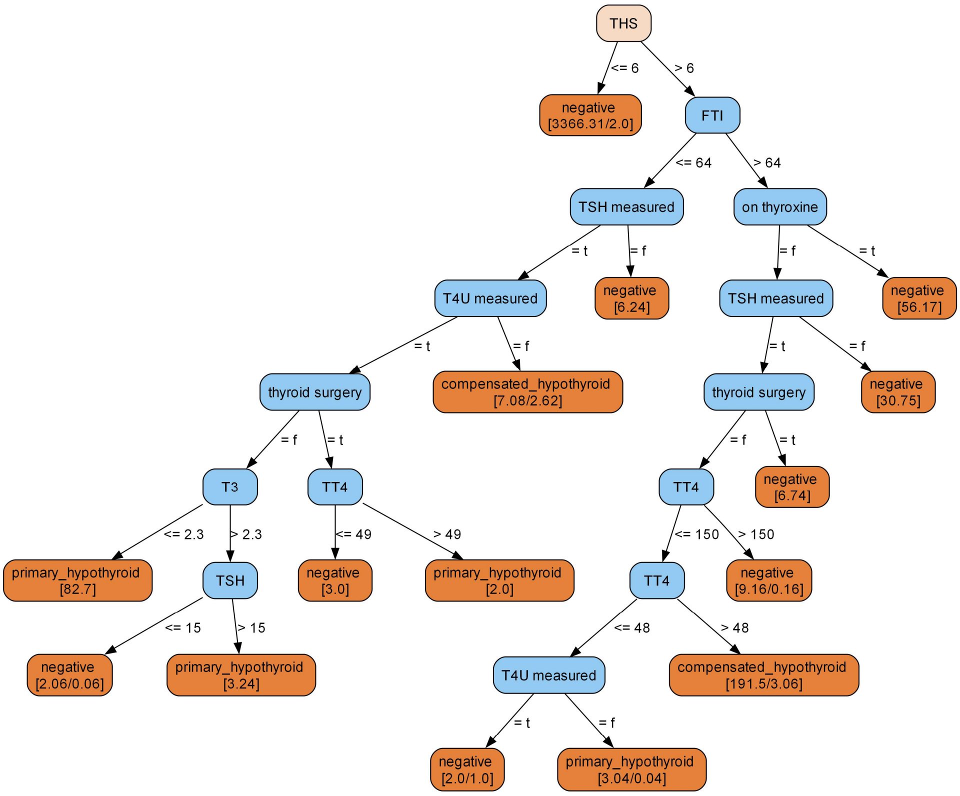


Figure 7. The base decision tree for the hypothyroid dataset.

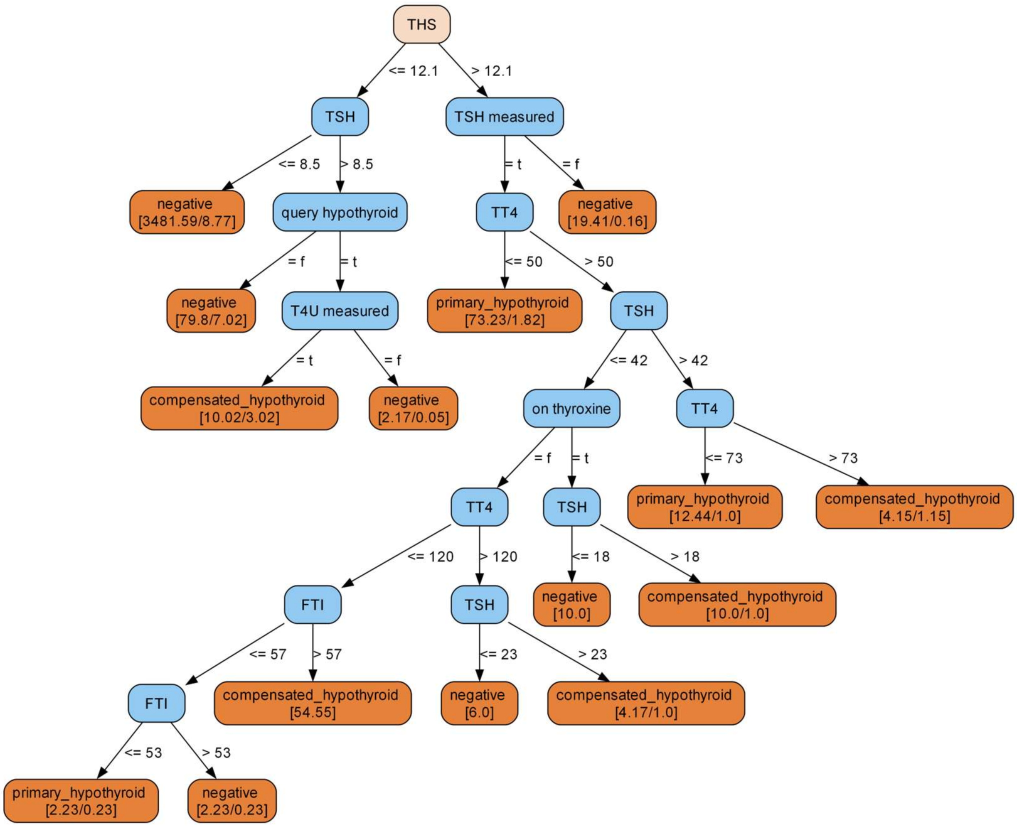


Figure 8. The decision tree for the hypothyroid dataset with the NaiveBayes classes model.

4. Discussion

This study utilized three distinct medical databases to explore the adaptation of the decision-making process in white-box models through knowledge infusion from black-box models. The results from each dataset illustrated how the decision-making processes could be altered to enhance transparency and performance.

In the breast cancer dataset, significant improvements in classification accuracy were achieved for the "recurrence-events" class through knowledge infusion, increasing the correctly classified instances by 16 or 18.8%. However, this was offset by a 12.4% decrease in correctly classified instances for the "no-recurrence-events" class. The differences between the outcomes of the black-box and the modified white-box models were minimal for the "no-recurrence-events" but showed a slight decline for "recurrence-events." The integration of the "inv-nodes" parameter markedly altered the visualization of decision-making processes, emphasizing the shift in the underlying mechanisms.

The diabetes dataset provided interesting insights during the cross-validation phase, where the black-box model was preferred, whereas the full dataset favored the white-box model. The knowledge-infused white-box model, though underperforming compared to the baseline white-box model on the complete dataset, showed improvements over the initial white-box model. Minor deviations of 1% in both the "tested_negative" and "tested_positive" classes were observed from the baseline black-box model, and a complete transformation in the decision-making process was evidenced by the changes in the decision trees.

For the hypothyroid dataset, the white-box algorithm unexpectedly outperformed all black-box counterparts, challenging the initial assumption of superior black-box performance. The knowledge transfer resulted in negligible deviations between the infused white-box and black-box models, demonstrating effective knowledge exchange as reflected by slight discrepancies in class values and notable changes in decision tree structures.

These findings underscore the potential of knowledge infusion in enhancing the explanatory depth and trustworthiness of decision-support systems in medicine—a field where the clarity and reliability of decision making are critical. The process of infusing knowledge from black-box to white-box models proves particularly valuable in handling imbalanced datasets where conventional white-box classifiers may yield unsatisfactory results. Notably, this methodology holds relevance even when the performance of black-box and white-box classifiers is comparable. However, it is crucial to rigorously assess the explanation layers of white-box models post-infusion to prevent misinterpretations in clinical decision making. The addition of black-box-infused visualization layers could offer novel insights, extending beyond traditional white-box explanations. The correlation observed between the decision-making processes of the black-box models and the infused white-box models indicates successful translation of decision-making elements into the white-box frameworks. This approach, by leveraging white-box algorithm visualizations to elucidate black-box decision making, provides medical professionals with the opportunity to assess the reasoning behind AI decisions critically. This insight not only highlights potential flaws in the decision-making process but also enables the re-evaluation and adoption of new and more effective solutions within clinical settings.

Nevertheless, this study identifies critical limitations. Knowledge infusion is a technique used to improve the interpretability of a white-box model by integrating insights from a black-box model. This process involves transferring patterns, decision-making cues, and other relevant information from the black-box model to the white-box model. The goal is to make the white-box model's decisions more understandable while maintaining or improving its accuracy. However, if a black-box model already performs perfectly, this means that for the given dataset and task, the model is making decisions with the maximum possible accuracy. This leaves no room for improvement in terms of accuracy or performance. Consequently, there is no variance or error in the model's predictions that can be analyzed and used to enhance the white-box model. Variance in this context refers to the differences or errors between the predicted outcomes by the model and the actual outcomes, which are used as learning points to improve another model's performance or interpretability. Furthermore, this methodology currently applies only to structured data, limiting its broader applicability.

Future research should aim to automate the testing of knowledge-infused models to systematically assess their impact, enabling practitioners to pinpoint the most effective interpretation layers for use with high-performing black-box models. While this methodology is showcased in medical contexts, its principles may be applicable in other high-stakes environments, complementing existing XAI solutions to boost transparency. Additionally, alternative methods that mimic decision-making processes, such as instance weighting, should be explored for potential synergies with this approach, broadening the scope and effectiveness of model transparency and interpretability in critical applications. Further testing is essential to fully elaborate on the reasoning behind and the feasibility of knowledge transfer in the presented methodology. Exploring alternative directions for knowledge transfer, such as reversing the flow between models or establishing a chain of transfers, could provide deeper insights into the inner workings and enhance the effectiveness of the proposed method. Such investigations could help refine this approach, ensuring a more robust framework for understanding and applying knowledge transfer in complex decision-making environments.

Beyond this knowledge-infusion approach, other notable XAI methodologies contribute to enhancing the transparency and interpretability of machine learning models, particularly in complex fields like medicine. For instance, Counterfactual Explanations, which highlight minimal changes needed to alter a model's decision, offer practical insights into decision boundaries [40]. Similarly, Feature Importance methods, exemplified by techniques like Gradient-based Feature Attribution (Grad-CAM) and Integrated Gradients, provide a granular understanding of input features that influence model outputs [41,42]. These methods complement the presented knowledge-infusion approach by enabling users to understand model reasoning from different perspectives, which can be critical in clinical settings where understanding the why behind a model's decision can be as important as the decision itself [43]. Combining these techniques with the presented approach could potentially yield a more robust framework for XAI in medicine, accommodating a broader range of clinical applications and ensuring that AI-driven decisions are both transparent and trustworthy [44].

5. Conclusions

This study explored the integration of black-box and white-box models in a medical context to enhance the interpretability of machine learning decisions without sacrificing performance. This research presented a novel approach to knowledge transfer between disparate models, focusing on the use of decision trees as a mechanism for introducing transparency into the inherently opaque processes of black-box algorithms. The findings from the three medical datasets, breast cancer, diabetes, and thyroid disease, underscore the viability of the presented methodology. By infusing decision-making insights from black-box models into white-box models, we effectively aligned the decision processes of white-box models with their black-box counterparts. The adapted white-box models exhibited significant correlations with the original black-box models, highlighting the effectiveness of this approach in maintaining accuracy while enhancing transparency. This was particularly evident in the models' ability to handle minority class instances more effectively, a crucial aspect in medical applications where accurate classification can be lifesaving.

The practical implications of this study are substantial. In medical settings where decision clarity is paramount, this approach not only facilitates a deeper understanding of AI-driven decisions but also builds trust among medical professionals and patients. Furthermore, the method allows for a nuanced exploration of the decision-making criteria employed in black-box models, providing insights that could lead to more informed and ethical AI implementations in healthcare. Despite the successes reported, the journey towards fully interpretable AI systems in high-stakes environments remains challenging. Future work will need to focus on refining these methods, possibly integrating more advanced forms of explainable AI techniques, and testing the approach in other complex datasets. Moreover, the broader application of these methods across different domains could be explored to validate the adaptability and robustness of the proposed methodology.

In conclusion, this study marks a significant step toward bridging the gap between the high performance of black-box models and the necessity for transparency in critical decision-making processes. By enhancing the interpretability of AI without undermining its effectiveness, we move closer to a future where AI can be trusted and utilized to its fullest potential in sensitive fields such as medicine.

Author Contributions: Conceptualization, B.Ž. and P.K.; methodology, B.Ž.; software, B.Ž.; validation, J.Z., H.B.V. and P.K.; formal analysis, P.K. and J.Z.; investigation, B.Ž., J.Z., H.B.V. and P.K.; writing—original draft preparation, B.Ž.; writing—review and editing, P.K. and H.B.V.; visualization, B.Ž.; supervision, J.Z. and P.K.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: This research received no external funding.

References

1. Bisaso, K.R.; Anguzu, G.T.; Karungi, S.A.; Kiragga, A.; Castelnuovo, B. A Survey of Machine Learning Applications in HIV Clinical Research and Care. *Comput. Biol. Med.* **2017**, *91*, 366–371. [CrossRef] [PubMed]
2. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; van der Laak, J.A.W.M.; van Ginneken, B.; Sánchez, C.I. A Survey on Deep Learning in Medical Image Analysis. *Med. Image Anal.* **2017**, *42*, 60–88. [CrossRef] [PubMed]
3. Bhattacharya, S.; Reddy Maddikunta, P.K.; Pham, Q.-V.; Gadekallu, T.R.; Krishnan S, S.R.; Chowdhary, C.L.; Alazab, M.; Jalil Piran, M. Deep Learning and Medical Image Processing for Coronavirus (COVID-19) Pandemic: A Survey. *Sustain. Cities Soc.* **2021**, *65*, 102589. [CrossRef] [PubMed]
4. Saraswat, D.; Bhattacharya, P.; Verma, A.; Prasad, V.K.; Tanwar, S.; Sharma, G.; Bokoro, P.N.; Sharma, R. Explainable AI for Healthcare 5.0: Opportunities and Challenges. *IEEE Access* **2022**, *10*, 84486–84517. [CrossRef]
5. Sharma, A.; Rani, R. *Machine Learning Applications in Anti-Cancer Drug Discovery. EAI/Springer Innovations in Communication and Computing*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 163–179. [CrossRef]
6. Lee, C.H.; Yoon, H.-J. Medical Big Data: Promise and Challenges. *Kidney Res. Clin. Pract.* **2017**, *36*, 3–11. [CrossRef] [PubMed]
7. Obermeyer, Z.; Emanuel, E.J. Predicting the Future—Big Data, Machine Learning, and Clinical Medicine. *N. Engl. J. Med.* **2016**, *375*, 1216. [CrossRef] [PubMed]
8. Libbrecht, M.W.; Noble, W.S. Machine Learning Applications in Genetics and Genomics. *Nat. Rev. Genet.* **2015**, *16*, 321–332. [CrossRef] [PubMed]
9. Suzuki, K. Overview of Deep Learning in Medical Imaging. *Radiol. Phys. Technol.* **2017**, *10*, 257–273. [CrossRef]
10. Wang, S.; Summers, R.M. Machine Learning and Radiology. *Med. Image Anal.* **2012**, *16*, 933–951. [CrossRef]
11. Xiao, Y.; Wu, J.; Lin, Z.; Zhao, X. A Deep Learning-Based Multi-Model Ensemble Method for Cancer Prediction. *Comput. Methods Programs Biomed.* **2018**, *153*, 1–9. [CrossRef]
12. Narula, S.; Shameer, K.; Salem Omar, A.M.; Dudley, J.T.; Sengupta, P.P. Machine-Learning Algorithms to Automate Morphological and Functional Assessments in 2D Echocardiography. *J. Am. Coll. Cardiol.* **2016**, *68*, 2287–2295. [CrossRef]
13. Senders, J.T.; Staples, P.C.; Karhade, A.V.; Zaki, M.M.; Gormley, W.B.; Broekman, M.L.D.; Smith, T.R.; Arnaout, O. Machine Learning and Neurosurgical Outcome Prediction: A Systematic Review. *World Neurosurg.* **2018**, *109*, 476–486.e1. [CrossRef] [PubMed]
14. Durán, J.M.; Jongsma, K.R. Who Is Afraid of Black Box Algorithms? On the Epistemological and Ethical Basis of Trust in Medical AI. *J. Med. Ethics* **2021**, *47*, 329–335. [CrossRef] [PubMed]
15. Bunge, M. A General Black Box Theory. *Philos. Sci.* **1963**, *30*, 346–358. [CrossRef]
16. Loyola-González, O. Black-Box vs. White-Box: Understanding Their Advantages and Weaknesses From a Practical Point of View. *IEEE Access* **2019**, *7*, 154096–154113. [CrossRef]
17. Rudin, C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [CrossRef] [PubMed]
18. Pintelas, E.; Livieris, I.E.; Pintelas, P. A Grey-Box Ensemble Model Exploiting Black-Box Accuracy and White-Box Intrinsic Interpretability. *Algorithms* **2020**, *13*, 17. [CrossRef]

19. Dresner, K.; Stone, P. A Multiagent Approach to Autonomous Intersection Management. *J. Artif. Intell. Res.* **2008**, *31*, 591–656. [CrossRef]
20. Caesar, H.; Bankiti, V.; Lang, A.H.; Vora, S.; Liong, V.E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; Beijbom, O. nuScenes: A Multimodal Dataset for Autonomous Driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11618–11628.
21. Tjoa, E.; Guan, C. A Survey on Explainable Artificial Intelligence (Xai): Toward Medical Xai. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 4793–4813. [CrossRef]
22. Antoniadi, A.M.; Du, Y.; Guendouz, Y.; Wei, L.; Mazo, C.; Becker, B.A.; Mooney, C. Current Challenges and Future Opportunities for XAI in Machine Learning-Based Clinical Decision Support Systems: A Systematic Review. *Appl. Sci.* **2021**, *11*, 5088. [CrossRef]
23. Van der Velden, B.H.; Kuijf, H.J.; Gilhuijs, K.G.; Viergever, M.A. Explainable Artificial Intelligence (XAI) in Deep Learning-Based Medical Image Analysis. *Med. Image Anal.* **2022**, *79*, 102470. [CrossRef]
24. Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
25. Sun, Y.-L.; Zhang, D.-L. Machine Learning Techniques for Screening and Diagnosis of Diabetes: A Survey. *Teh. Vjesn.* **2019**, *26*, 872–880. [CrossRef]
26. Fryer, D.; Strümke, I.; Nguyen, H. Shapley Values for Feature Selection: The Good, the Bad, and the Axioms. *IEEE Access* **2021**, *9*, 144352–144360. [CrossRef]
27. Dieber, J.; Kirrane, S. Why Model Why? Assessing the Strengths and Limitations of LIME. *arXiv* **2020**, arXiv:2012.00093.
28. Gunning, D.; Aha, D.W. DARPA's Explainable Artificial Intelligence Program. *AI Mag.* **2019**, *40*, 44–58. [CrossRef]
29. The Precise4Q Consortium; Amann, J.; Blasimme, A.; Vayena, E.; Frey, D.; Madai, V.I. Explainability for Artificial Intelligence in Healthcare: A Multidisciplinary Perspective. *BMC Med. Inf. Decis. Mak.* **2020**, *20*, 310. [CrossRef] [PubMed]
30. Žlahtič, B.; Završnik, J.; Blažun Vošner, H.; Kokol, P.; Šuran, D.; Završnik, T. Agile Machine Learning Model Development Using Data Canyons in Medicine: A Step towards Explainable Artificial Intelligence and Flexible Expert-Based Model Improvement. *Appl. Sci.* **2023**, *13*, 8329. [CrossRef]
31. Kerr, A.D.; Scharp, K. The End of Vagueness: Technological Epistemicism, Surveillance Capitalism, and Explainable Artificial Intelligence. *Minds Mach.* **2022**, *32*, 585–611. [CrossRef]
32. Zolanvari, M.; Yang, Z.; Khan, K.; Jain, R.; Meskin, N. TRUST XAI: Model-Agnostic Explanations for AI With a Case Study on IIoT Security. *IEEE Internet Things J.* **2023**, *10*, 2967–2978. [CrossRef]
33. Cabitza, F.; Campagner, A.; Datteri, E. To Err Is (Only) Human. Reflections on How to Move from Accuracy to Trust for Medical AI. In *Exploring Innovation in a Digital World*; Ceci, F., Prencipe, A., Spagnoletti, P., Eds.; Lecture Notes in Information Systems and Organisation; Springer International Publishing: Cham, Switzerland, 2021; Volume 51, pp. 36–49. ISBN 978-3-030-87841-2.
34. Kingsford, C.; Salzberg, S.L. What Are Decision Trees? *Nat. Biotechnol.* **2008**, *26*, 1011–1013. [CrossRef] [PubMed]
35. Podgorelec, V.; Kokol, P.; Stiglic, B.; Rozman, I. Decision Trees: An Overview and Their Use in Medicine. *J. Med. Syst.* **2002**, *26*, 445–463. [CrossRef] [PubMed]
36. Attwal, K.P.S.; Dhiman, A.S. Exploring Data Mining Tool-Weka and Using Weka to Build and Evaluate Predictive Models. *Adv. Appl. Math. Sci.* **2020**, *19*, 451–469.
37. Witten, I.H.; Frank, E.; Hall, M.A.; Pal, C.J. *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed.; Elsevier/Morgan Kaufmann: San Francisco, CA, USA, 2016; ISBN 978-0-12-804291-5.
38. Java | Oracle. Available online: https://www.java.com (accessed on 7 February 2024).
39. Arnold, K.; Gosling, J.; Holmes, D. *The Java Programming Language*, 4th ed.; Addison-Wesley: Upper Saddle River, NJ, USA, 2006; ISBN 978-0-321-34980-4.
40. Dai, X.; Keane, M.T.; Shalloo, L.; Ruelle, E.; Byrne, R.M.J. Counterfactual Explanations for Prediction and Diagnosis in XAI. In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, Oxford, UK, 19–21 May 2021; ACM: Oxford, UK, 2022; pp. 215–226.
41. Ancona, M.; Ceolini, E.; Öztireli, C.; Gross, M. Towards Better Understanding of Gradient-Based Attribution Methods for Deep Neural Networks. *arXiv* **2017**, arXiv:1711.06104.
42. Sayres, R.; Taly, A.; Rahimy, E.; Blumer, K.; Coz, D.; Hammel, N.; Krause, J.; Narayanaswamy, A.; Rastegar, Z.; Wu, D.; et al. Using a Deep Learning Algorithm and Integrated Gradients Explanation to Assist Grading for Diabetic Retinopathy. *Ophthalmology* **2019**, *126*, 552–564. [CrossRef] [PubMed]
43. Sahoh, B.; Choksuriwong, A. The Role of Explainable Artificial Intelligence in High-Stakes Decision-Making Systems: A Systematic Review. *J. Ambient Intell. Humaniz. Comput.* **2023**, *14*, 7827–7843. [CrossRef]
44. Chamola, V.; Hassija, V.; Sulthana, A.R.; Ghosh, D.; Dhingra, D.; Sikdar, B. A Review of Trustworthy and Explainable Artificial Intelligence (XAI). *IEEE Access* **2023**, *11*, 78994–79015. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.