Abstract: This paper proposes a novel ensemble approach, integrating Artificial Neural Networks (ANNs), Symbolic Regression (SR), and Decision Trees (DTs), to predict fish mortality caused by infectious diseases. The intensifying global burden of fish diseases threatens the sustainability of aquatic ecosystems and the aquaculture industry, necessitating sophisticated modelling strategies for effective disease management and control. The proposed approach capitalizes on the non-linear data modelling strength of ANNs, the explanatory power of SR, and the decision-making efficiency of DTs, offering both predictive accuracy and interpretable insights. The architecture of the proposed ensemble method is developed in two stages. In the intermediate stage, an ANN is employed to learn the complex, non-linear interactions between various biological and environmental factors impacting fish health. Additionally, SR is applied to produce a symbolic equation that effectively maps the input variables to fish mortality rates. In the final stage, a DT model is included to enhance prediction performance by capturing decision rules from the data. This hybrid approach offers superior prediction performance while also revealing meaningful biological/environmental relationships that can guide preventive and reactive interventions in the management of fish health. We evaluate the developed models using extensive real-world datasets acquired from two large Greek fish-farming units, which encompass representative disease types. The results demonstrate that our ensemble approach significantly outperforms traditional standalone models developed in our recent previous work, achieving enhanced predictive accuracy, robustness, and interpretability. Overall, this research has far-reaching implications for improving disease predictions, facilitating optimal decision-making in aquaculture management, and contributing to the sustainability of global fish stocks.

Keywords: fish death estimation; aquaculture; machine learning; Artificial Neural Networks; Symbolic Regression; Decision Trees; ensemble modelling

1. Introduction

The sustainability of global aquatic ecosystems and the aquaculture industry is increasingly under threat due to the surge in infectious diseases among fish populations [1]. These diseases not only affect the health and survival of fish species but also pose significant economic challenges to fisheries and aquaculture [2]. The accurate prediction and understanding of fish mortality due to infectious diseases are, therefore, crucial in managing and mitigating these threats effectively.

Traditional modelling strategies, while offering some insights, often struggle to capture the complex interactions between various biological and environmental factors influencing fish health. Artificial Neural Networks (ANNs) have emerged as a powerful tool in this...
context due to their ability to learn non-linear patterns and interactions in large, complex datasets, acting as universal function approximators [3–7]. Yet, despite their predictive prowess, ANNs often act as “black boxes”, offering little insight into the relationships between the variables used in predictions. On the other hand, Symbolic Regression (SR) offers interpretable models that can elucidate the underlying structure of data by producing symbolic equations [8,9] (The higher interpretability of SR has made it a favourable approach used for extracting the underlying laws of physical systems from experimental data [8,9]). In a similar vein, Decision Trees (DTs), rooted in foundational works by Breiman et al. [10], provide a methodical decomposition of the data space into decision nodes and leaves, enabling the straightforward interpretation of the decision paths, which enhances transparency and aids in understanding the feature influence on predictions. Unfortunately, however, the high interpretability of SR and DTs often comes at the expense of prediction performance, particularly when dealing with complex, non-linear relationships.

Against this background, in this work, we introduce an ensemble approach that integrates the distinct advantages of ANNs, SR, and DTs. The proposed method not only leverages the high predictive accuracy of ANNs but also combines the interpretability of SR, with the structured decision-making capabilities of DTs. Developed in two stages (intermediate and final), our ensemble method achieves robust results with respect to fish mortality due to infectious diseases and also reveals meaningful relationships among key influencing factors.

This article extends our recent previous work published in [11], where standalone ANN and SR models were developed using real-world datasets from the Greek aquaculture industry, specifically tailored to fish diseases caused by the bacteria 

*Pasteurella*, *Vibrio Harveyi* and *Myxobacteria*. Building on [11], we present the architecture and evaluation of our innovative ensemble models, each developed for a specific disease. Utilizing extensive real-world datasets acquired from two large Greek fish-farming units, which cover diseases caused by 

*Pasteurella*, *Vibrio Harveyi*, *Myxobacteria*, and *Viral Nervous Necrosis* (a disease not studied in [11] due to lack of data), we demonstrate the superiority of our approach over the standalone ANN and SR models developed in [11].

Overall, the present research contributes to the field of aquaculture management by providing a powerful tool for predicting disease outcomes, thus facilitating better decision-making and contributing to the sustainability of global fish stocks. The insights gained from the study also open up new avenues for future research, aimed at refining the proposed models for more specific diseases and contexts, further bolstering their applicability and efficacy in real-world scenarios.

The remainder of this article is structured as follows: The next section explores notable relevant literature. Section 3 discusses the profile of the available data, on which the implemented ensemble models shall be trained and evaluated. Thereafter, Section 4 introduces the architecture of the approach followed and discusses the technical aspects of the ANN, SR, and DT models. Section 5 presents and comments on the obtained results. The article closes with a brief conclusion section, which summarizes our contribution and highlights interesting avenues for future investigation.

2. Related Work

The individual application of ANNs, SR models, and DTs constitutes a common ground in various domains; however, the conjunctive use of these Machine Learning (ML) models—as implemented in this work—is still relatively scarce in the literature. In this section, we present a representative collection of such conjunctive implementations.

The first work that we discuss is [12], in which an ANN-based architecture for SR is proposed, which is integrated with other Deep Learning architectures, such that the whole system can be trained end-to-end through backpropagation. The authors demonstrated the performance of their model on several substantially different tasks and showed that it can extrapolate quite well outside of the training dataset compared to a standard ANN-based architecture.
In [13], a novel hybrid model that combines SR and a deep Multilayer Perceptron (MLP) for one-month-ahead photovoltaic-power forecasting is proposed. The system was evaluated in a case-study analysis, using a real Australian weather dataset, where the employed input features were solar irradiation and the historical photovoltaic power data. The overall study revealed favourable properties of the proposed hybrid SR-MLP approach. In a similar vein, the authors of [14] used ANNs and SR to predict the energy consumption in public buildings at the University of Granada. Their study concluded that there are no significant differences in accuracy considering both techniques in the problems addressed.

In [15], a collection of optimization studies was performed for helically finned tubes in heat exchangers, in the context of which a comparison of ANNs and SR-based correlations was conducted, whereas, in [16], an ANN and an SR model were conjunctively employed and revealed a completely new criterion for predicting glass-forming ability, a property that represents the difficulty of forming metallic glasses.

Furthermore, the authors of [17] developed and evaluated an ANN and an SR model to predict instantaneous exhaust emission in transient conditions of a diesel engine, fuelled with animal fat in different proportions. Lastly, in [18], a framework based on SR, coupled with physics-informed neural networks, was proposed (physics-informed neural networks constitute a special type of ANNs that can embed the knowledge of any physical laws that govern a given dataset in the learning process and can be described by partial differential equations [19,20]). The framework was tailored to uncover the unknown parts of non-linear equations of motion directly from data.

As far as the use of DTs along with ANNs and/or SR is concerned, the authors of [21] introduced PS-Tree, an advanced ML algorithm that combines DTs with SR to handle non-linear, piecewise models efficiently. This method enhances traditional DTs by incorporating adaptive and evolutionary strategies to optimize feature construction and partition schemes dynamically. The success of PS-Tree in outperforming other state-of-the-art algorithms in a plethora of datasets underlines its effectiveness in handling complex modelling tasks. The use of SR to enhance DTs has also been explored in [22], where a novel technique that employs SR to enrich DTs splits was presented. The proposed method synergistically combines SR and DTs in order to discover short, compact trees, with much richer class decision boundaries. In [23], the integration of DTs and ANNs in learning-to-rank tasks for personal search applications is explored. The study introduces an ensemble method to leverage the strengths of both models, aiming to improve ranking effectiveness and deployment efficiency in personal search systems. Lastly, Pagliarini et al., in [24], discuss the challenges and solutions associated with multivariate time-series classification, particularly emphasizing the balance between interpretability and accuracy in ML models. Motivated by these remarks, the authors propose a hybrid model that integrates the strengths of both ANNs and temporal DTs. The proposed model captures temporal patterns through ANNs while maintaining decision-making transparency through temporal DTs.

Having discussed an indicative collection of relevant literature, we close this section by stressing that, to the best of our knowledge, the conjunctive use of ANN, SR, and DT models for estimating fish mortality in aquaculture is absent. This highlights the innovative aspect of our approach in integrating these diverse methodologies to enhance predictive accuracy and model reliability.

3. Data Profile

The available data for the study were obtained from two large fish-farming units based in Greece and cover a recent period of almost five years (that is, from 1 January 2018 to 11 October 2022). The data include daily records concerning the following quantities:

- Number of fish deaths due to a specific infectious disease, and total number of fish deaths (due to all causes);
- Water temperature;
- Administered amount of food;
- Administered amount of medicated food;
• Administered antibiotics doses;
• Administered vaccination doses.

The infectious diseases examined herein are those for which there were a sufficient amount of data available to effectively train the developed ML models; those diseases are the diseases caused by the bacteria *Pasteurella, Vibrio Harveyi* and *Myxobacteria*, as well as *Viral Nervous Necrosis* (VNN). We note that Viral Nervous Necrosis was not studied in our previous work [11], due to the lack of data at the time the research was conducted. Moreover, the data for Pasteurella, Vibrio Harveyi and Myxobacteria used herein have been augmented relative to [11].

Against this background, Figure 1 depicts the time-series of the daily fish deaths due to the examined diseases, as well as the time-series of the daily total fish deaths (due to all causes), as recorded in the available data, i.e., during the recent period of almost five years. (The time-series of the daily fish deaths due to the examined diseases, as depicted in Figure 1, present discontinuities, since the depicted fish deaths concern multiple (rather than a single) fish cages of a fish-farming unit. This means that a time-series, as shown in Figure 1, firstly, represents all consecutive daily fish deaths recorded in fish cage *A*, which then represents all consecutive daily fish deaths recorded in fish cage *B*, and so forth. We retain this illustration of time-series for a better presentation. On the other hand, the time-series of the daily total fish deaths does not present such discontinuities, since the recorded total fish deaths refer to all fish cages of a fish-farming unit. Note, lastly, that the values on the axes of all time-series presented herein are not shown at the request of the fish-farming units, due to confidentiality agreements. Therefore, the reported time-series are designed to illustrate changes in fish deaths, rather than absolute values. It is important to emphasize that the absence of axis values does not impede the deployment and validation of the introduced ensemble models, which is the aim of the study.) It is evident that fish deaths due to Pasteurella and VNN are more frequent than fish deaths due to Vibrio Harveyi and Myxobacteria, which suddenly appear in rare time-intervals. As far as the daily total fish deaths are concerned, the two high peaks appearing in Figure 1 are noteworthy. We note also that the volume of the available data differs for each considered disease. In particular, the number of samples/records referring to fish deaths due to Pasteurella and VNN is significantly greater than the number of samples/records referring to fish deaths due to Vibrio Harveyi and Myxobacteria.

![Figure 1](image-url)
Figure 1. Time-series of the daily fish deaths due to the examined diseases, as well as of the daily total fish deaths (due to all causes), during the recent period of almost five years.

4. Materials and Methods

As illustrated in Figure 2a, for each infectious disease under consideration D, a trained ensemble model takes as input a dataset of the last N days of a particular fish cage (i.e., data for the present day t₀, data for the day before today t₋₁, data for the day before yesterday t₋₂, and so forth) and produces as an output a real number that represents the estimate for the fish deaths due to the disease D, on the j-th forthcoming day t₊j. The input data corresponding to a particular day i concern the real-world available data discussed in Section 3; hence, the input data of day i include the fish deaths due to D on the day i, the water temperature on day i, the amount of food given on day i, the administered medicated food of day i, and the antibiotics and vaccination doses administered on the day i. By combining the output of each trained ensemble model (which pertains to a specific disease), an estimate for the fish deaths of the j-th forthcoming day, due to all considered infectious diseases, is produced. For our case study, it is assumed that each ensemble model receives as input the data of a fish cage for the last 10 days (thus, N = 10) and produces as an output an estimate for the 3rd forthcoming day (thus, j = 3). These specific numbers were chosen in consultation with the experts of the fish-farming units.

The ensemble models are constructed in two stages. The first, or intermediate, stage incorporates both an ANN and an SR model. The second, or final, stage not only includes the ANN and SR models, but also adds a DT model. The architecture diagram of each one of these two stages is illustrated in Figure 2. As shown in Figure 2b, an intermediate stage ensemble model integrates an ANN and an SR model to estimate the number of fish deaths; on the other hand, as suggested in Figure 2c, a final stage ensemble model also incorporates a DT model to further support the estimation. In any case, the synthesis of the results of the ANN, SR, and DT models is implemented by means of a weighted voting regressor, which combines the distinct ML models and computes a weighted average of their predictions. The selection of weights for the weighted voting regressor—specifically, 0.5, 0.3, and 0.2 for the ANN, SR, and DT models, respectively—has been determined through a series of trials, systematically evaluating various combinations to optimize
the balance between the models based on their individual performance and consistency. This trial-based approach ensures that each model’s contribution to the final prediction is proportional to its predictive accuracy and reliability, leading to more robust and precise ensemble outcomes.

![Block diagram of the proposed ensemble approach](image)

**Figure 2.** Block diagram of the proposed ensemble approach (a) and architecture diagrams of an intermediate stage (b) and a final stage (c) ensemble model.

It is noteworthy that the above approach to estimating fish deaths can also be applied as a binary classification approach, which does not capture the absolute number of estimated fish deaths but only whether fish deaths are expected or not. This can be carried out as follows. If the output of a trained ensemble model, deployed for a disease \( D \), is a positive real number (or a positive real number greater than a certain pre-specified threshold), then we assume that fish deaths, due to \( D \), are indeed expected on the \( j \)-th forthcoming day. If the output of that ensemble model is zero (or a number close to zero), then we assume that no fish deaths, due to \( D \), are expected on the \( j \)-th forthcoming day. In the context of this binary classification method, the generated output will now be a collection of binary numbers (0 and 1), rather than a collection of real numbers.

In addition to the ensemble models trained to estimate fish deaths due to infectious diseases, we also developed an ensemble model trained for the estimation of the total number of fish deaths (due to all causes). Similar to the previous ensemble models, this
model is built in two stages: intermediate and final. The model’s input consists of the total number of fish deaths recorded on each one of the last 4 days, and its output is a real number representing the estimated total number of fish deaths of the next day. Again, these specific numbers were chosen in consultation with the experts of the fish-farming units. It should be stressed that all the developed ensemble models are evaluated in a validation dataset that is distinct from the training dataset, and its size is about 40% of the size of the training dataset.

4.1. Artificial Neural Networks

Let us firstly discuss the architecture and parameters of the developed ANNs. All the ANNs considered have identical structures, so we shall focus on a single ANN. Accordingly, the alluded ANN is a conventional Multilayer Perceptron (MLP), which consists of four layers, namely, one input layer, two fully connected (dense) hidden layers, and one output layer (cf. Figure 3). The input layer of the network is a passive layer with no learnable parameters, which merely receives the data of each sample of the dataset. Each fully connected hidden layer has 130 neurons (units) and employs Rectified Linear Unit (ReLU) activation functions. The output layer contains a single linear neuron, which produces a real number that represents the estimation of fish deaths due to a particular infectious disease. The ANN is compiled using a Mean Squared Error (MSE) loss function and the well-known Adam optimization algorithm, which is an extension of the stochastic gradient descent [25] (It is noted that the ANNs developed in [11] are also compiled using a Mean Squared Error loss function, rather than using a Mean Absolute Error, as stated in the text of [11], due to a typographic error). For tackling the overfitting problem, the dropout regularization technique was used where necessary. By implementing dropout, the selected neurons of the ANN are randomly ignored (“dropped out”) during the training phase, and, as a consequence, no weight updates are applied to those ignored neurons [26].

![Figure 3. Architecture of an ANN in charge of a particular disease.](image-url)

4.2. Symbolic-Regression Models

Symbolic Regression (SR) is a method of interpretable Machine Learning that reduces data to mathematical equations. More precisely, SR is a type of regression analysis that searches the space of mathematical expressions (symbolic expressions) to find the model that best fits a given dataset in terms of both accuracy and simplicity. Contrary to conventional regression techniques that seek to optimize parameters for a predefined structure of a regression model, SR does not impose prior assumptions and instead infers both the
structure of a regression model and its parameters directly from the data. This, in turn, entails that SR is not affected by human bias or gaps in the domain knowledge, since it does not require a priori specification of a regression model (which is perhaps mathematically comprehensive from a human perspective). Yet a disadvantage of SR is the fact that it takes considerable time to find a regression model that appropriately fits the dataset, as the corresponding search space is huge. Nevertheless, most SR algorithms prevent a combinatorial explosion by applying evolutionary algorithms that iteratively improve—over generations—the mathematical expression that best fits the available data.

Such an evolutionary algorithm, based on the fundamental concepts of Darwinian evolution, builds an initial population of naïve random formulas, by randomly combining mathematical building blocks, such as mathematical operators (i.e., $+, -, \times, \div$), analytic functions (e.g., $\sqrt{}$, $\cos$, $\sin$, $\exp$, $\log$), constants, and state variables. Each successive generation of formulas, then, evolves from the preceding one, by selecting the fittest individuals from the population to undergo genetic operations. Eventually, the algorithm takes a series of totally random formulas, untrained and unaware of any given target function and makes them breed, mutate and evolve their way towards a formula that best fits the data.

In our study, the development of SR models was carried out using Python’s `gplearn` tool, which extends the `scikit-learn` ML library to perform Genetic Programming for Symbolic Regression [27]. Table 1 lists the hyper-parameters set up for the `gplearn` tool. The explanation of each hyper-parameter is as follows: Population Size is the number of individuals (mathematical formulas) in each generation, Generations is the maximum number of generations, Metric is the measure of an individual’s fitness (herein MAE), and Stopping Criteria expresses the fitness value at which the evolution procedure terminates. Function Set contains the mathematical functions used when building and evolving generations. Crossover Probability controls the crossover method according to which genetic material between individuals is mixed. Sub-tree Mutation Probability, Hoist Mutation Probability, and Point Mutation Probability control the respective mutation operations. Lastly, the Parsimony Coefficient is a constant that penalizes large individuals by adjusting their fitness (MAE) to be less favourable for selection—this penalty helps in producing less computationally costly individuals which are, at the same time, more understandable.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population Size</td>
<td>5000</td>
</tr>
<tr>
<td>Generations</td>
<td>20</td>
</tr>
<tr>
<td>Metric</td>
<td>Mean Absolute Error (MAE)</td>
</tr>
<tr>
<td>Stopping Criteria</td>
<td>0.01 (fish deaths)</td>
</tr>
<tr>
<td>Function Set</td>
<td>add, sub, mul, div, $\sqrt{}$, $\cos$, $\sin$, log</td>
</tr>
<tr>
<td>Crossover Probability</td>
<td>0.7</td>
</tr>
<tr>
<td>Sub-tree Mutation Probability</td>
<td>0.1</td>
</tr>
<tr>
<td>Hoist Mutation Probability</td>
<td>0.05</td>
</tr>
<tr>
<td>Point Mutation Probability</td>
<td>0.1</td>
</tr>
<tr>
<td>Parsimony Coefficient</td>
<td>0.05-0.1</td>
</tr>
</tbody>
</table>

4.3. Decision Tree Models

Decision Tree (DT) models are integral components of many ensemble methods due to their inherent interpretability, ease of implementation, and capability to handle non-linear data relationships. In contrast to linear models, DTs partition the data space into regions defined by rules based on feature thresholds, allowing them to capture complex patterns through a simple, hierarchical decision structure. This makes them particularly useful in
diverse settings, ranging from risk assessment and medical diagnosis to financial analysis, where the decisions need to be both accurate and understandable.

In each ensemble model developed herein, we incorporate a DT regressor, chosen for its ability to model non-linear relationships and interactions between features through successive binary decisions. The architecture and parameters of the utilized DT were meticulously selected to optimize both performance and generalizability, as detailed below:

- **Criterion and Splitting Strategy:** The DT utilizes the “squared error” criterion to assess the quality of splits within the tree. This measure focuses on minimizing the variance of the target values within each node, aiming to produce homogenized predictions. It uses the “best” splitter strategy, which exhaustively considers all possible splits across all features, selecting the one that provides the most significant reduction in squared error, thereby optimizing each decision path for the greatest predictive accuracy.

- **Tree Depth and Complexity:** The maximum depth of the tree is constrained to 10 levels. This limitation not only prevents the model from overfitting, by curbing the complexity of the decision paths, but also maintains computational efficiency and interpretability. While deeper trees might model the training data with high fidelity, they often fail to generalize well; hence, controlling the depth helps balance bias and variance.

- **Leaf Nodes and Samples:** To further prevent overfitting, each leaf/terminal node in the tree must contain at least five samples. This parameter helps avoid creating overly specific rules, thus reducing model variance, without a significant increase in bias. The tree also ensures that a node splits only if it contains at least two samples, allowing the tree to expand sufficiently, but avoiding insignificant divisions.

- **Node Purity and Randomness:** There are no constraints on the maximum number of leaf nodes, and no minimum impurity decrease threshold is enforced, suggesting that the tree’s expansion is primarily influenced by its depth and the minimum requirements in leaf samples. No monotonic constraints are applied, offering the flexibility to capture both increasing and decreasing trends in feature relationships.

Through the aforementioned settings, the DT of each ensemble model is designed to achieve an optimal balance between prediction accuracy and computational efficiency.

5. Results and Discussion

To demonstrate the derived results, Table 2 is first reported, which, for each infectious disease under consideration, presents the Mean Absolute Error (MAE) obtained by the corresponding final stage ensemble model in the validation dataset. For the sake of comparison, Table 2 also presents the corresponding MAEs of the intermediate stage ensemble models, as well as of the ANN and SR models developed in our previous work [11] (recall that VNN was not studied in [11]). It is evident that, for each of the diseases Pasteurella, Vibrio Harveyi and Myxobacteria, the corresponding ensemble models achieved a lower MAE than the corresponding standalone ANN and SR models of [11].

Specifically, for Pasteurella, the final stage ensemble model reduced the MAE to 52.18 from 60 (ANN) and 65.1 (SR), representing improvements of approximately 13% and 20%, respectively. The intermediate stage ensemble model also showed improvement over the standalone models of [11], with an MAE of 58.55, though not as significant as the final stage model. In the case of Vibrio Harveyi, the final stage ensemble model achieved an MAE of 13.13, compared to 18 (ANN) and 20 (SR), marking reductions of about 27% and 34%, respectively. The intermediate stage ensemble model’s MAE was 13.24, showing significant improvement as well, but slightly higher than the final stage model, indicating a minor difference. For Myxobacteria, the final stage ensemble model’s MAE of 2.95 is lower than the ANN’s MAE of 3 and significantly better than the SR’s MAE of 9.13, showcasing a minor improvement over ANN and a substantial 68% reduction over SR. The intermediate stage ensemble model had an MAE of 2.98, indicating consistent performance improvement over the standalone models of [11], but with a marginal difference compared to the final stage model. Overall, the comparison between intermediate and final stage ensemble models reveals that, while both stages offer significant improvements over the standalone
ML models of [11], the final stage ensemble models consistently achieve slightly better MAE values than the corresponding intermediate stage models.

Table 2. Mean Absolute Error (MAE) in the validation dataset of the (two-staged) ensemble models, as well as of the standalone ANN and SR models developed in [11], for each infectious disease. Note that VNN was not studied in [11].

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Pasteurella</td>
<td>52.18</td>
<td>58.55</td>
<td>60</td>
<td>65.1</td>
</tr>
<tr>
<td>Vibrio Harveyi</td>
<td>13.13</td>
<td>13.24</td>
<td>18</td>
<td>20</td>
</tr>
<tr>
<td>Myxobacteria</td>
<td>2.95</td>
<td>2.98</td>
<td>3</td>
<td>9.13</td>
</tr>
<tr>
<td>Viral Nervous Necrosis (VNN)</td>
<td>1.27</td>
<td>1.54</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

Thereafter, for each one of the examined diseases, Figure 4 depicts the estimation of the intermediate and final stage ensemble models in the validation dataset for the number of fish deaths on the 3rd forthcoming day. Firstly, it is evident that the estimations of the intermediate and final stage ensemble models for Pasteurella and VNN display no significant visual differences; in contrast, the intermediate and final stage ensemble models for Vibrio Harveyi and Myxobacteria demonstrate noticeable visual variations. In general, the ensemble models for Pasteurella, Myxobacteria, and VNN demonstrate a close alignment between real and estimated data points, indicating a high degree of model accuracy. Particularly noteworthy are the models for Myxobacteria, which nearly perfectly capture the actual spike, suggesting exceptional predictive precision. However, the models for Vibrio Harveyi show a significant discrepancy during peak events, where the models underestimate the severity of the impact.

Let us now turn to the symbolic knowledge of the SR models encompassed in the trained (intermediate and final stage) ensemble models. Accordingly, we subsequently present the mathematical expressions encoded into each SR model developed for the estimation of fish deaths caused by Pasteurella, Vibrio Harveyi, Myxobacteria and VNN (Due to the randomness characterizing the construction of the SR models, not all runs converge into exact mathematical expressions. The mathematical expressions shown herein are indicative. The units of measurement for all variables involved in the equations are identical to those of the available data, discussed in Section 3).

\[
\begin{align*}
\text{Deaths}_{\text{Past}}_{+3} &= \sqrt{\frac{\text{Deaths}_{\text{Past}0}}{\text{Deaths}_\text{Temp}_{-3}}} + \frac{\text{Deaths}_{\text{Past}0}}{\text{VACCINE}_{1-6} \cdot \text{VACCINE}_{1-3} \cdot \text{Temp}_{-3}} \\
\text{Deaths}_{\text{Vib}}_{+3} &= \text{MED}_{\text{FOOD}}_{I-7} \\
\text{Deaths}_{\text{Myx}}_{+3} &= \text{MED}_{\text{FOOD}}_{II-7} \\
\text{Deaths}_{\text{VNN}}_{+3} &= \text{VACCINE}_{I0} + \text{MED}_{\text{FOOD}}_{III-4}
\end{align*}
\]

Some comments on the above mathematical equations are presented in order.
Figure 4. Estimation (in the validation dataset) of the intermediate-stage (a) and final-stage (b) ensemble models for the number of fish deaths on the 3rd forthcoming day, due to the considered infectious diseases.
• The first equation asserts that the estimated number of fish deaths due to Pasteurella on the 3rd forthcoming day (Deaths_Past+3) is a function of the number of fish deaths due to Pasteurella recorded today (Deaths_Past0) and six days ago (Deaths_Past−6), of the water temperature recorded three days ago (Temp−3), as well as of the amount of the vaccines “VACCINE_I” and “VACCINE_II” administered six and three days ago (VACCINE_I−6 & VACCINE_I−3).

• The second expression is a very simple equation, asserting that the estimated number of fish deaths due to Vibrio Harveyi on the 3rd forthcoming day (Deaths_Vib+3) is a function of the amount of the medicated food “MED_FOOD_I” administered seven days ago (MED_FOOD_I−7).

• The third expression is again a very simple equation, asserting that the estimated number of fish deaths due to Myxobacteria on the 3rd forthcoming day (Deaths_Myx+3) is a function of the amount of the medicated food “MED_FOOD_II” administered seven days ago (MED_FOOD_II−7).

• Lastly, the fourth equation states that the estimated number of fish deaths due to VNN on the 3rd forthcoming day (Deaths_VNN+3) is a function of the vaccine “VACCINE_I” administered today (VACCINE_I0), as well as of the medicated food “MED_FOOD_III” administered four days ago (MED_FOOD_III−4).

It should be evident that the mathematical equations of the SR models encompassed into the ensemble models, essentially, provide fish farmers with a precise and quantifiable method for describing the development of fish mortalities resulting from infectious diseases.

Thereafter, the Decision Trees of the DT models encompassed into the trained final stage ensemble models are concisely presented in Figure 5. Note that, for an efficient presentation, only the root node and the next two levels of each Decision Tree are shown. The depicted Decision Trees exemplify the capability of the ensemble models to tailor their predictive mechanisms to the peculiarities of each pathogen, namely, Pasteurella, Vibrio Harveyi, Myxobacteria, and VNN. The Decision Trees shown in Figure 5 not only furnish a transparent decision-making process but also reveal critical biological and environmental thresholds that significantly influence mortality outcomes. For instance, factors like water temperature and administered amounts of food and medicated food, which are prominently featured in the Decision Tree concerning Pasteurella, are identified as pivotal in the disease’s propagation. Such insights are indispensable for aquaculture operators, as they highlight specific conditions under which preventive measures could be most effectively deployed.

We close this section with the presentation of the results concerning the total number of fish deaths (due to all causes), during the examined time period. Accordingly, Figure 6 depicts the estimation of the developed final stage ensemble model in the validation dataset, for the total number of fish deaths on the next day. For that case, the MAE of the final stage ensemble model is 1850, which is thus lower than the MAE of the standalone ANN and SR models developed in [11], which was 1900 and 2070, respectively. This represents an improvement of approximately 2.6% over the ANN, and a more substantial 10.6% reduction compared to the SR model. The intermediate stage ensemble model achieved an MAE of 1880, indicating a 1% improvement over the ANN and a 9.2% reduction compared to the SR model of [11]. While both ensemble stages demonstrate significant enhancements over the standalone models of [11], the final stage ensemble model achieves slightly better results than the intermediate-stage model, reflecting a consistent trend of incremental improvement in the final stage.
Figure 5. A concise presentation of the Decision Trees of the DT models of the final stage ensemble models. The variables shown in the nodes of the trees are as follows: Deaths\(_i\): number of fish deaths due to a disease recorded \(i\) days ago, Temp\(_i\): water temperature recorded \(i\) days ago, Food\(_i\): amount of food administered \(i\) days ago, Medicated_Food\(_j\): amount of the medicated food “\(j\)” administered \(i\) days ago. Nodes that do not correspond to a variable are leaf/terminal nodes.
Figure 6. Estimation (in the validation dataset) of the final stage ensemble model for the total number of fish deaths on the next day.

Lastly, a concise presentation of the Decision Tree of the corresponding DT model is illustrated in Figure 7.

Figure 7. A concise presentation of the Decision Tree of the DT model encompassed into the respective final-stage ensemble model. Variable $Deaths_i$ denotes the total number of fish deaths recorded $i$ days ago.

Lastly, the mathematical expression encoded into the respective SR model is stated subsequently.

$$D_{+1} = 0.746 \cdot D_{-3} + \sqrt{\frac{D_{-1} \cdot D_{-2} - \frac{1}{2} \sqrt{D_{-2}} - 0.85 \cdot D_{-3} - 0.85 \cdot \frac{3}{2} \sqrt{D_{-3}} + 0.104}{D_{-3}}}$$

The derived mathematical expression asserts that the total number of fish deaths on the next day ($D_{+1}$) is a (rather complex) function of the total number of fish deaths recorded on each one of the last four days ($D_0, D_{-1}, D_{-2},$ and $D_{-3}$).

6. Conclusions

This study presented a novel ensemble approach, combining Artificial Neural Networks (ANNs), Symbolic Regression (SR), and Decision Trees (DTs) for predicting fish mortality due to infectious diseases. The implemented models, built in two stages (intermediate and final), were designed to address the need for accurate prediction tools, which also provide interpretable insights into the interactions between influential factors. Our findings revealed that these hybrid models offer a notable improvement over the standalone ANN and SR models developed in our earlier work [11], as they achieved superior predictive accuracy across all the examined disease types. Specifically, the final stage ensemble models achieved an MAE of 52.18, 13.24, 2.95 and 1.27 for the diseases of Pasteurella, Vibrio Harveyi, Myxobacteria and Viral Nervous Necrosis, respectively. These values are lower...
than those achieved by the standalone ANN and SR models of [11] in the corresponding diseases. Additionally, the intermediate-stage ensemble models also showed significant improvements, with MAEs of 58.55, 13.24, 2.98, and 1.54 for Pasteurella, Vibrio Harveyi, Myxobacteria and Viral Nervous Necrosis, respectively, demonstrating that both stages of the ensemble method enhance predictive accuracy compared to standalone models.

By integrating the strengths of ANNs, SR, and DTs, our approach offers a powerful tool for aquaculture management, facilitating better decision-making and helping ensure the sustainability of global fish stocks. The capacity to predict and understand fish mortality due to infectious diseases could greatly aid in the development of preventive measures and effective treatments, which are of utmost importance in the rapidly growing field of aquaculture.

Future work is to be devoted to the evaluation of more advanced ANN-based models, more sophisticated SR models (such as the “AI Feynman” [9]), and further exploration of the synergistic potential of DTs within the ensembles. Moreover, future research will aim to refine the developed ensemble models for more specific diseases and contexts, further enhancing their real-world applicability. As our knowledge of fish diseases expands, the models can be updated and improved to incorporate new findings, making them a versatile and evolving tool in fish health management.

Author Contributions: Conceptualization, I.H. and G.S.; Methodology, T.A. and I.H.; Software, T.A.; Validation, T.A.; Formal analysis, T.A. and I.H.; Investigation, T.A.; Resources, G.S.; Writing—original draft, T.A.; Writing—review & editing, I.H. and G.S.; Visualization, G.S.; Supervision, I.H. 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: The datasets presented in this article are not readily available due to confidentiality agreements with the fish-farming units from which the data were obtained. Requests to access the datasets should be directed to Georgios Spiliopoulos (g.spiliopoulos@kefish.gr).

Conflicts of Interest: Author Georgios Spiliopoulos was employed by the company Kefalonia Fisheries S.A. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References


