International Journal of
Translational Medicine

MDPI

Article

From Measurements to Patients: Data Aggregation in Supervised Classification of X-Ray Diffraction Datasets

Alexander Alekseev [1,2], Keith Rogers [3,4], Lev Mourokh [3,5,*] and Pavel Lazarev [1,3]

1 Matur UK Ltd., 5 New Street Square, London EC4A 3TW, UK; aalexeev@matur.co.uk (A.A.); plazarev@matur.co.uk (P.L.)
2 Department of Physics and Technology, Karaganda Buketov University, Karaganda 100028, Kazakhstan
3 EosDx, Inc., 15211 Vanowen Street, Suite 209, Los Angeles, CA 91405, USA
4 Shrivenham Campus, Cranfield University, Swindon SN6 8LA, UK; k.d.rogers@cranfield.ac.uk
5 Physics Department, Queens College, City University of New York, 65-30 Kissena Blvd., Flushing, NY 11367, USA
* Correspondence: lev.murokh@qc.cuny.edu

Abstract

Background/Objectives: Machine learning approaches are widely used in modern medical diagnostics, including cancer detection. The results can be significantly improved by aggregating individual measurements, and appropriate aggregation methods should be established. Methods: We applied various measurement aggregation strategies both before and after machine learning modeling to two datasets of X-ray diffraction images: human breast biopsy samples and canine claw samples. Two classifiers, Random Forest and Logistic Regression, were used to determine classification metrics: the area under the receiver operating characteristic curve (ROC-AUC) and balanced accuracy. Results: We found that all aggregation types improve classification metrics, with aggregation after modeling yielding better performance. Depending on the dataset and approach, either classifier can produce better results. For human breast samples, Random Forest with the logit aggregation strategy provides an ROC-AUC exceeding 0.9. For the canine dataset, both Random Forest with the logit aggregation strategy and Logistic Regression with the median of cancer probabilities achieve an ROC-AUC of about 0.85. Conclusions: We examined several simple, straightforward aggregation methods for patient diagnosis based on multiple measurements per patient and achieved significant improvements in classification metrics.

Keywords: machine learning; aggregation; supervised classification; X-ray diffraction

Check for updates

Academic Editor: Simone Brogi

Received: 12 April 2026
Revised: 10 May 2026
Accepted: 13 May 2026
Published: 15 May 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The use of supervised machine learning in biomedical diagnostics has increased substantially in recent years, particularly in applications involving medical imaging, molecular assays, and multi-omics data [1–4]. In many clinical settings, diagnostic decisions are based not on a single measurement, but on multiple, heterogeneous observations collected at different levels of biological or technical hierarchy. Typical examples include repeated measurements of the same specimen, multiple samples obtained from a single patient, or results from different diagnostic methods, such as imaging, molecular assays, and clinical markers.

The effectiveness of the resulting diagnostic model can be improved by aggregating various intermediate outputs, accounting for their structure and statistics. Developing efficient procedures for such aggregation remains an important methodological challenge

in clinical machine learning. Existing approaches include those based on statistical parameters [5], various ensembling methods (voting, bagging, stacking) [6–8], and more sophisticated neural-network-based techniques of the Multi-Instance Learning (MIL) family [9–13].

In this work, we examine a simple case in which aggregation is performed from a set of identical measurements at different locations on a single sample and over the set of samples belonging to a single patient. This is analogous to existing situations, such as the analysis of radiographic and histological images, in which each sample typically contains components of both normal and pathological tissues. Since the development of digital methods and automated diagnostics, several approaches have been explored to identify tissue as cancerous, even when most of the image appears "normal". Processing usually involves subdividing images into multiple tessellating regions of interest ("tiles"), extracting features from each tile, and subsequently aggregating the information from each tile to form a single, slide-level classification [14–16].

Here, we examine aggregation methods only at two levels of hierarchy, from measurements to patients, ignoring the intermediate sample level. Despite the apparent simplicity of the problem statement, the number of possible aggregation options is substantial and requires substantial computational resources, especially given the statistical approach used in our work.

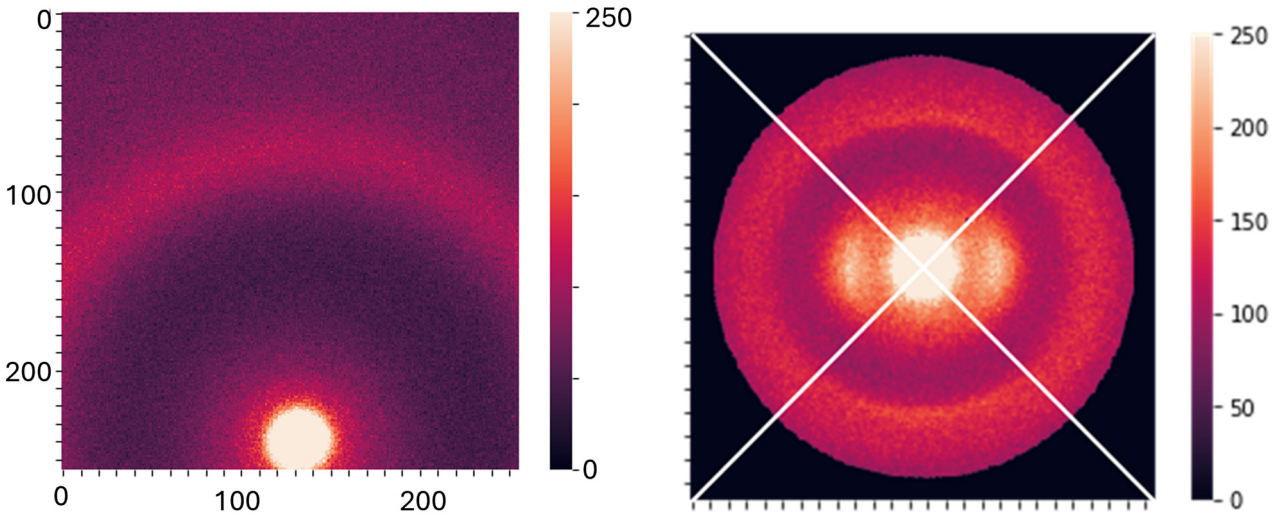
For the first time, we have explored methods of aggregate processing for X-ray scatter imaging. Recently, X-ray scattering of biological tissues has been used as a detection tool to reveal novel structural biomarkers, i.e., components of the extracellular matrix altered by cancer [17–26]. We used two specific datasets of X-ray diffraction (XRD) images, one from human breast cancer biopsy samples [22–24] and another from canine nails, both cancerous and healthy [25,26]. In the first case, each patient provided two biopsy samples, and, depending on sample dimensions, measurements were performed at four, five, or nine locations. In the second case, the XRD patterns were obtained from three, four, or five claws per patient. We implemented two distinct approaches to the "measurement-to-patient" transition: aggregation preceding machine learning cancer/non-cancer binary modeling, and aggregation of the machine learning results. As the dataset sizes are insufficient to apply deep learning procedures, our approaches are simple and straightforward. For the first approach, we used various statistical characteristics and their combinations as objects for the classification. For the second approach, various aggregation techniques were employed. In both cases, we used and compared two classifiers, Random Forest and Logistic Regression. It should be emphasized that the goal of this work is not to achieve the best classification metrics on these specific datasets, but rather to determine general trends in aggregations of multi-level and multi-modal data. The advantages of these methods are their ease of implementation and the simplicity of interpreting results.

2. Materials and Methods

2.1. Description of the Datasets

In this work, we used XRD data from human breast cancer biopsy samples obtained from the Breast Cancer Now Biobank, UK, and from the nails of canines diagnosed with cancer. Measurements from control and healthy groups supplemented both datasets. The measurement procedure and image pre-processing steps have been previously described for human samples in [22,23] and for canine nails in [25,26]. For the biopsy samples, we used results from small-angle X-ray scattering (SAXS) and wide-angle X-ray scattering (WAXS), whereas for the nails, only the WAXS data are relevant. After pre-processing, Fourier coefficients were calculated for all XRD images. For the biopsy samples, coefficients were obtained from a 2D Fourier transform of the XRD pattern (for example, Figure 1a),

and only the first 10 orders were retained (304 coefficients). For the nails, the azimuthally averaged profiles were initially obtained in the horizontal and vertical directions (e.g., Figure 1b), and thereafter, 1D Fourier coefficients were calculated. We kept the real and imaginary parts for 20 coefficients of each profile.



(**a**) (**b**)

Figure 1. (**a**) Typical diffraction pattern of the human breast biopsy sample [23]; (**b**) Typical diffraction pattern of the canine nail [25] with the sectors indicating the regions of the azimuthal averaging for vertical and horizontal directions.

The hierarchical structure of the datasets is presented in Table 1. The canine dataset has only a single hierarchical level. In contrast, the biopsy dataset includes several levels related to transitions from measurements to samples, from samples to patients, and to the combination of SAXS and WAXS data. There are two biopsy samples per patient, and each sample was measured at 4, 5, or 9 locations, so the total number of measurements per patient ranges from 8 to 18. For canines, the number of nails per patient can be 3, 4, or 5.

Table 1. Structure of the datasets.

	Breast Cancer Data				**Canine Claw Data**	
	WAXS		**SAXS**			
Patients (no cancer/cancer)	282 (123/159)				249 (149/100)	
Samples	564				920	
Measurements	4067		4074		920	
Number of patients in the train set	200				170	
Number of patients with a given number of measurements	8 9 10 13 14 18	40 14 21 23 50 134	8 9 10 13 14 17 18	40 12 21 25 50 1 133	3 4 5	77 171 1

2.2. Aggregation Strategies

In this work, we explored two aggregation approaches: (i) prior to machine learning modeling, when the cancer probability was determined from the aggregated measurements, and (ii) post-machine learning modeling, when the cancer probabilities from individual measurements were aggregated. In both cases, the sample step is omitted, resulting in a direct transition from measurement to patient. In (i), for all Fourier coefficients belonging to a specific patient, the statistical characteristics with respect to the measurements were determined. We used 7 distribution characteristics: mean, standard deviation, minimum, maximum, skewness, kurtosis, and median. Correspondingly, for a given Fourier coefficient, the set of values related to a specific measurement is replaced by a set of values for the statistical characteristics. Next, classification was performed for each characteristic separately and for its combinations. The abbreviations and descriptions for classifications used in this work are shown in Table 2.

Table 2. Methods of probability aggregations used prior to machine learning modeling.

Abbreviation	Comments/Definitions
Waxs	A combination of mean, standard deviation, minimum, maximum, skewness, and kurtosis from WAXS data
Saxs	A combination of mean, standard deviation, minimum, maximum, skewness, and kurtosis from SAXS data
waxs_mean	Mean value of coefficients from WAXS data
waxs_std	Standard deviation for coefficients from WAXS data
waxs_min	Minimal value of coefficients from WAXS data
waxs_max	Maximal value of coefficients from WAXS data
waxs_median	Median for coefficients from WAXS data

For canines' nails, only WAXS data were used. Also, the number of measurements was insufficient to calculate kurtosis, so only 5 characteristics were used in "all_features".

In (ii), the classification was performed at the measurement level, and the resulting cancer probabilities were aggregated. We used various aggregation methods, with their abbreviations and descriptions presented in Table 3.

Table 3. Methods of probability aggregations used in post-machine learning modeling.

Abbreviation	Comments/Definitions
p_mean	Mean of probabilities per patient
p_median	Median of probabilities per patient
p_std	Standard deviation per patient
p_min	Min probability per patient
p_max	Max probability per patient
dist_mean	Mean distance from optimal decision threshold, d_i
w_p_mean	Mean of weighted probabilities, $d_i p_i$
c_w_p_mean	Mean of cubic-weighted probabilities, $d_i^3 p_i$
logit_mean	Mean of $\ln(p_i/(1-p_i))$

Here, the relative distance, d_i, from the optimal threshold is defined as

$$d_i = \begin{cases} (p_i - t)/t & \text{if} p_i < t \\ (p_i - t)/(1-t) & \text{if} p_i \geq t \end{cases} \quad (1)$$

where p_i is the probability of cancer for the i-th measurement and t is the optimal decision threshold, providing maximal balanced accuracy at the patient level. Values of d_i occur in the range $-1 \leq d_i \leq 1$, and were negative for healthy and positive for cancerous patients.

2.3. Classification Methods

Machine learning procedures were carried out using the Scikit-learn, NumPy, SciPy, and Pandas libraries. The diagnostic models were obtained from supervised learning. For comparison, we used two classifiers, Random Forest and Logistic Regression. Each model was optimized using hyperparameters to maximize the area under the receiver operating characteristic curve (ROC-AUC). The optimized models were then calibrated using Platt (sigmoid) scaling.

The general algorithm is as follows:

1. Only data at the level of measurements corresponding to (randomly selected) approximately 70% of patients were used to train the model. The remaining ~30% of patients were used for testing. Separating the training and testing sets by patient ensured no data leakage.
2. The selected training data was used to optimize the hyperparameters of the classifiers (Logistic Regression and Random Forest). Hyperparameter optimization was performed on a grid of values using GridSearchCV with 5-fold cross-validation and model efficiency scoring using ROC-AUC. The cross-validation was performed by grouping by patient IDs.
3. Calibration of probabilities using the Platt method was carried out using training data. Next, the optimal decision threshold was determined using the calibrated probabilities.
4. Subsequently, the metric was determined, taking into account the data aggregation method.
5. The entire procedure was repeated 100 times with different random_state values, responsible for selecting patients for training. Finally, 100 ROC-AUC and balanced accuracy values were obtained for each aggregation method.

The ROC-AUC dependencies on the number of splits for the two datasets and two classifiers are shown in Figure 2.

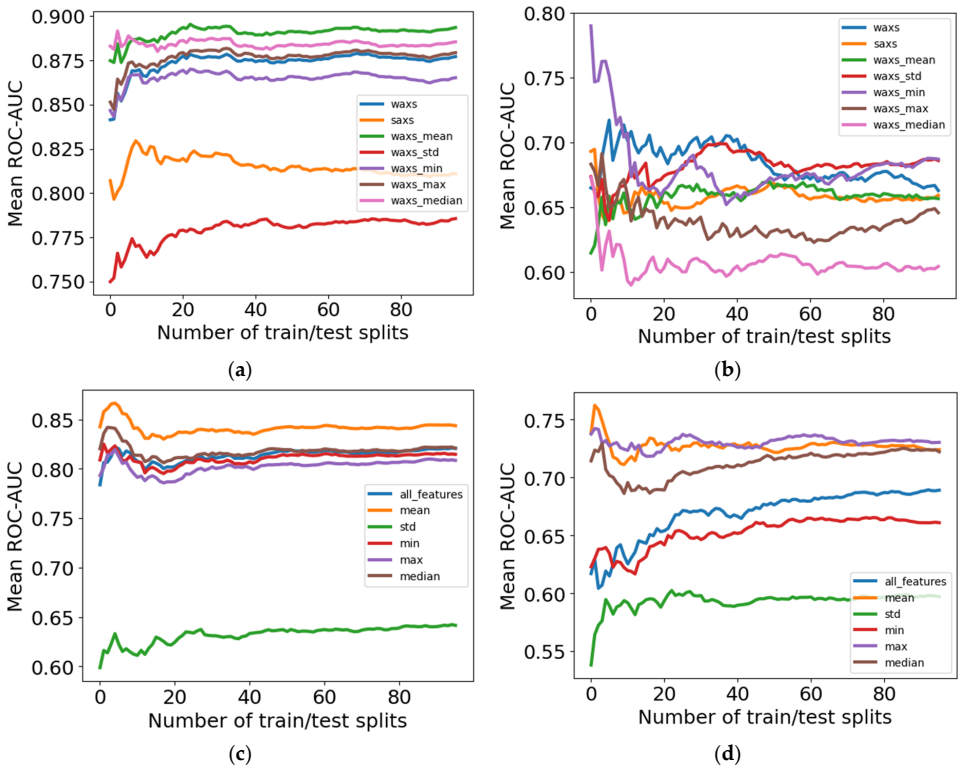


Figure 2. The averaged value of ROC-AUC as a function of the number of training/test splits. (**a**) Random Forest for the biopsy samples; (**b**) Logistic Regression for the biopsy samples; (**c**) Random Forest for the canines' nails; (**d**) Logistic Regression for the canines' nails.

In our analysis, we used 100 splits as there were no significant fluctuations of mean ROC-AUC estimates at this value, and the computation time remained reasonable. ROC-AUC and maximal achieved balanced accuracy were used as classification metrics. The 95% confidence interval for each metric was calculated by using the *t*-interval, as mean $\pm\, t \times$ SEM, where t is the positive critical value of the *t*-distribution curve, the standard error of the mean is SEM = $s/\sqrt{n}$, $n = 100$, and s is the sample standard deviation.

3. Results

3.1. Human Breast Samples

3.1.1. Aggregation Before Modeling

After the aggregation procedure described in Section 2.2, for both WAXS and SAXS data, the sets of values of specific Fourier coefficients were replaced by their statistical parameters (mean, standard deviation, minimum, maximum, skewness, kurtosis, and median). The classification can be performed using each of these parameters separately or in various combinations. The classification of the complete sets of parameters using Random Forest produced ROC-AUC values of 0.878 [0.871, 0.883] for WAXS and 0.811 [0.803, 0.82] for SAXS, where the numbers in the square brackets indicate the 95% confidence intervals. For this approach, the WAXS results are significantly greater than the SAXS results, and thus we will restrict our discussion to these within this section. The ROC-AUC and balanced accuracy are shown in Figure 3, based on the statistical characteristics of the WAXS Fourier coefficients. A Random Forest classifier was used for panels (a) and (b), whereas Logistic Regression was employed for panels (c) and (d). The error bar indicates the 95% confidence interval.

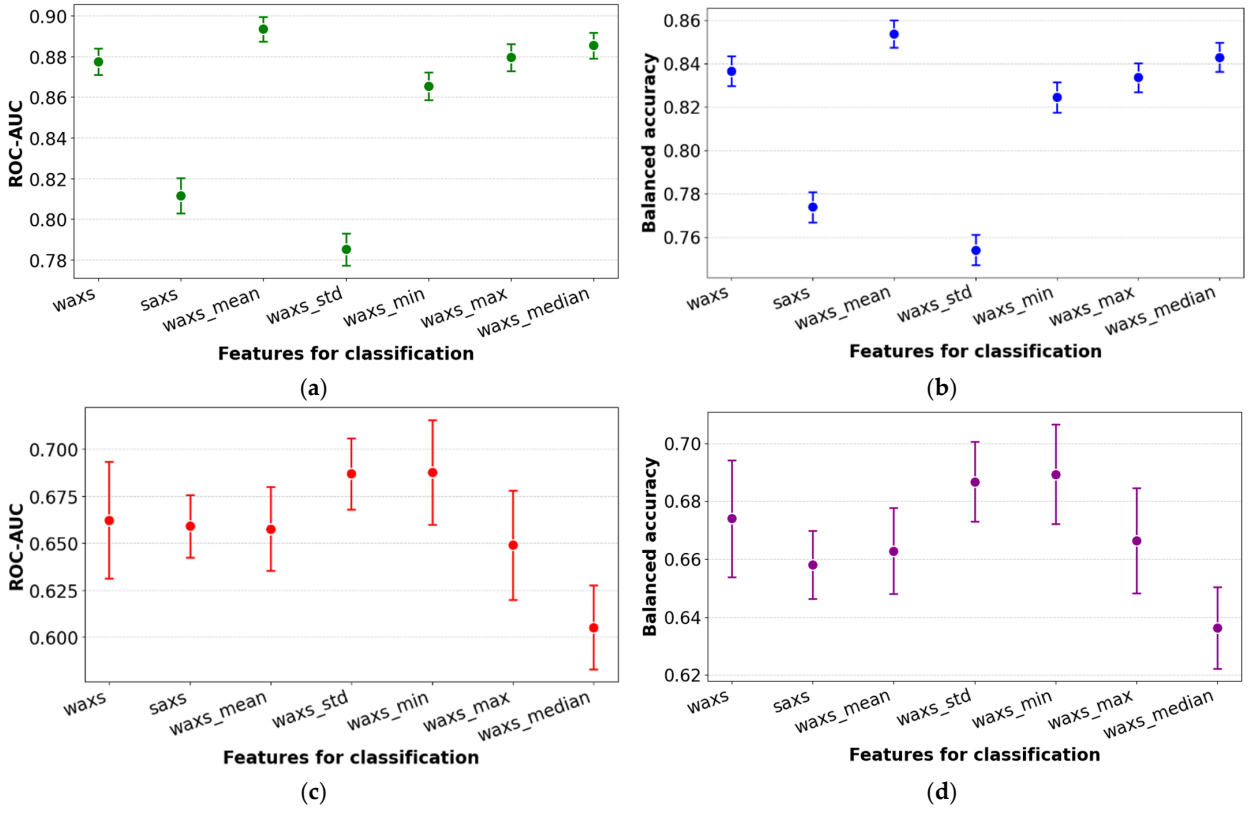


Figure 3. The classification results for human breast samples using various statistical parameters. (**a**) ROC-AUC for Random Forest classifier; (**b**) balanced accuracy for Random Forest classifier; (**c**) ROC-AUC for Logistic Regression classifier; (**d**) balanced accuracy for Logistic Regression classifier. The error bars indicate the 95% confidence interval.

Figure 3 shows that Random Forest significantly outperformed Logistic Regression, with improved metrics and a smaller confidence interval. Among the statistical metrics, the best results were obtained with mean values, producing an ROC-AUC of 0.894 [0.887, 0.899] and a balanced accuracy of 0.854 [0.847, 0.86]. While the Random Forest metrics for the standard deviation are poorer than those for the mean, they remain substantial, indicating that sample heterogeneity differs between healthy and cancerous patients. For Logistic Regression, they are comparable, with the best metrics provided by the minimum. The higher-order moments, skewness and kurtosis, do not produce significant classification performance, likely because they are noisy given the small number of measurements in the set.

3.1.2. Aggregation After Modeling

After classification at the measurement level, the aggregation strategies listed in Table 3 can be employed. The obtained ROC-AUC and balanced accuracy values are presented within Figures 4 and 5 for WAXS and SAXS, respectively. As in the previous section, for the results shown in panels (a) and (b), the Random Forest classifier was used, whereas for panels (c) and (d), Logistic Regression was employed. The results for measurements only, without aggregation, are also presented as "measur".

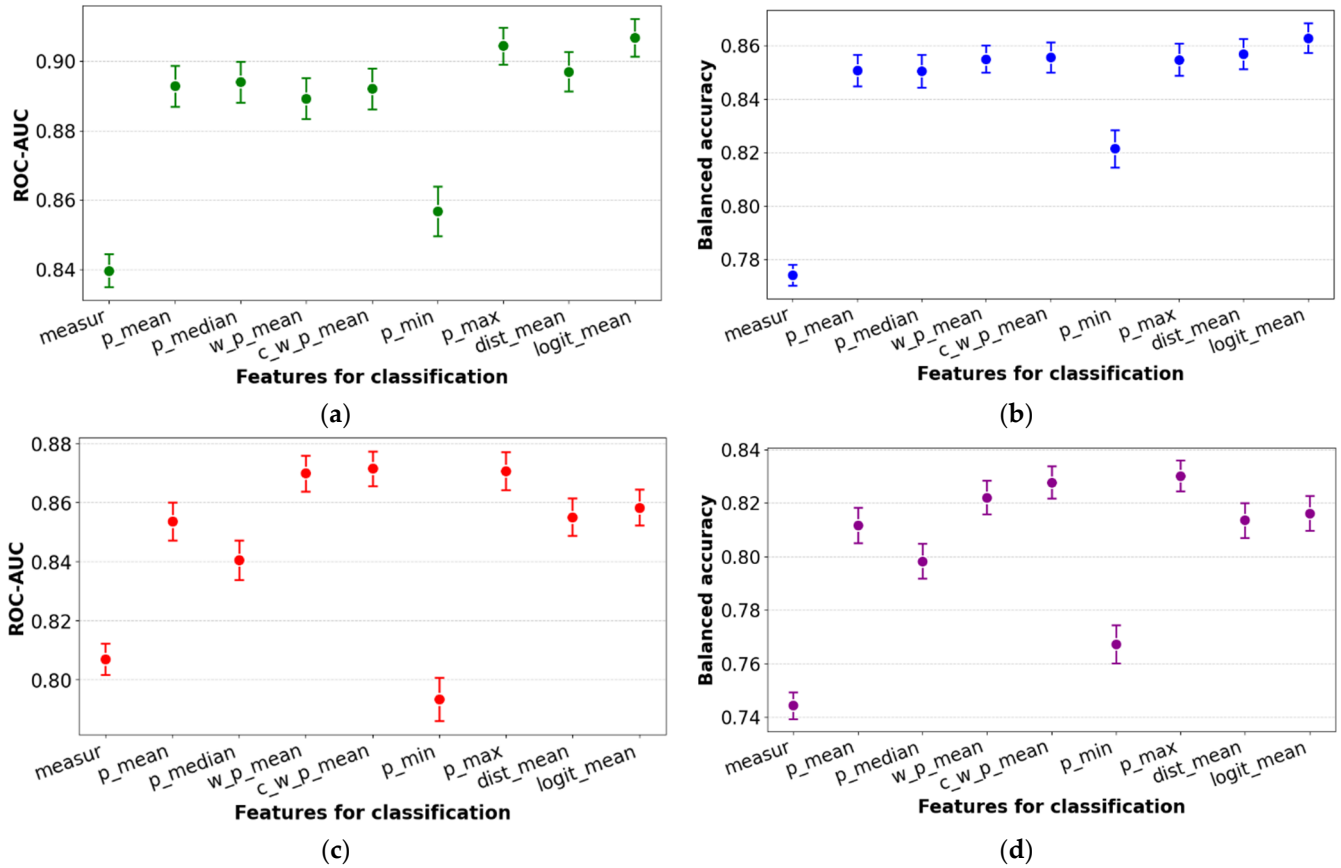


Figure 4. WAXS data: The classification results for human breast samples using various aggregation procedures listed in Table 3. (**a**) ROC-AUC for Random Forest classifier; (**b**) balanced accuracy for Random Forest classifier; (**c**) ROC-AUC for Logistic Regression classifier; (**d**) balanced accuracy for Logistic Regression classifier. The error bars indicate the 95% confidence interval.

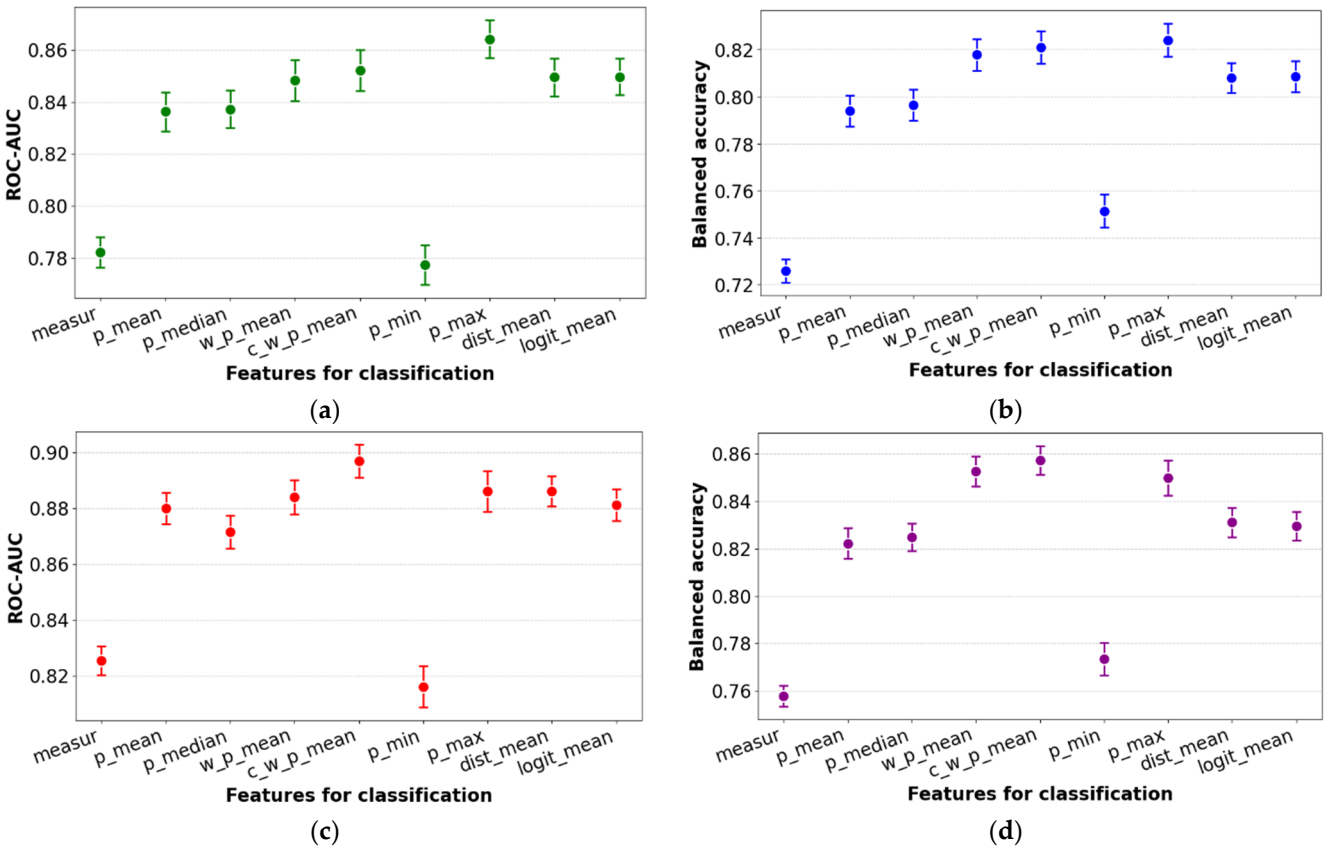


Figure 5. SAXS data: The classification results for human breast samples using various aggregation procedures listed in Table 3. (**a**) ROC-AUC for Random Forest classifier; (**b**) balanced accuracy for Random Forest classifier; (**c**) ROC-AUC for Logistic Regression classifier; (**d**) balanced accuracy for Logistic Regression classifier. The error bars indicate the 95% confidence interval.

It is evident from these figures that aggregation significantly improves classification performance. Random Forest outperforms Logistic Regression for WAXS patterns, but the situation is reversed for SAXS data. Metrics for cubic-weighted probabilities are consistently greater than those for probabilities or weighted probabilities. The best metrics were provided by the Random Forest classifier on WAXS data using the logit aggregation strategy, with an ROC-AUC of 0.907 and a balanced accuracy of 0.863.

We also determined the optimal number of positive measurements for a "hard prediction"; the diagnostic decision is made when the cancer probabilities for n measurements exceed the threshold. The ROC-AUC dependencies on n obtained by Random Forest and Logistic Regression are shown in Figure 6 for WAXS and SAXS.

Similar to the aggregation case, Random Forest is superior for WAXS, and Logistic Regression is better for SAXS. The optimal number of positive measurements for the hard prediction is five or six. However, even at the optimal numbers, the metrics are inferior to those for aggregation, making the latter procedure necessary for successful diagnostics.

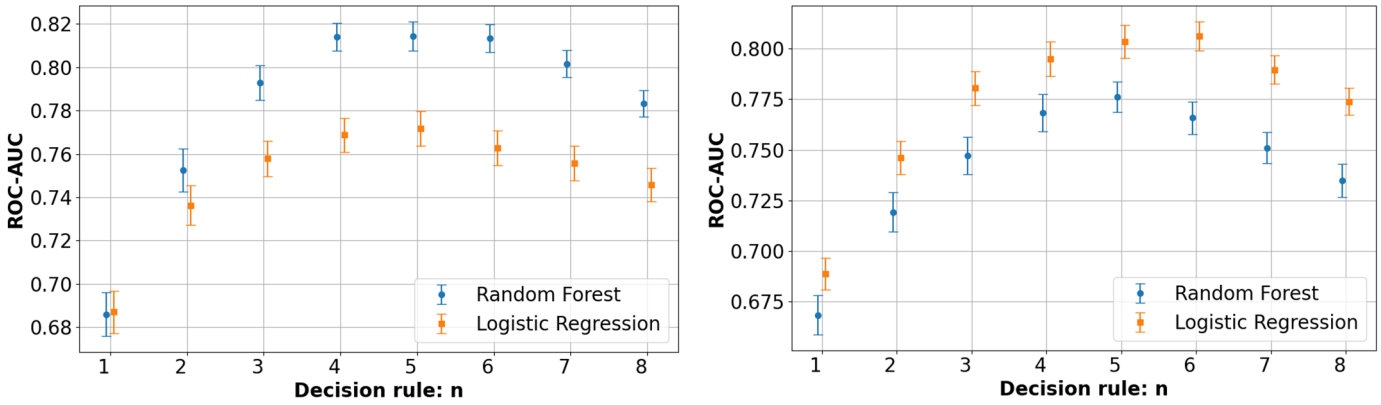


Figure 6. Hard prediction rule. (**a**) ROC-AUC for WAXS; (**b**) ROC-AUC for SAXS. The error bars indicate the 95% confidence interval.

3.2. Canine Claw Samples

3.2.1. Aggregation Before Modeling

The same measurement aggregation procedure was applied to the XRD dataset of canine nails. Accounting for all statistical characteristics provides an ROC-AUC of 0.821 [0.812, 0.83] for Random Forest and 0.69 [0.673, 0.705] for Logistic Regression. The balanced accuracy values are 0.774 [0.765, 0.783] and 0.679 [0.667, 0.69], respectively. The metrics for the separately taken statistical characteristics are shown in Figure 7.

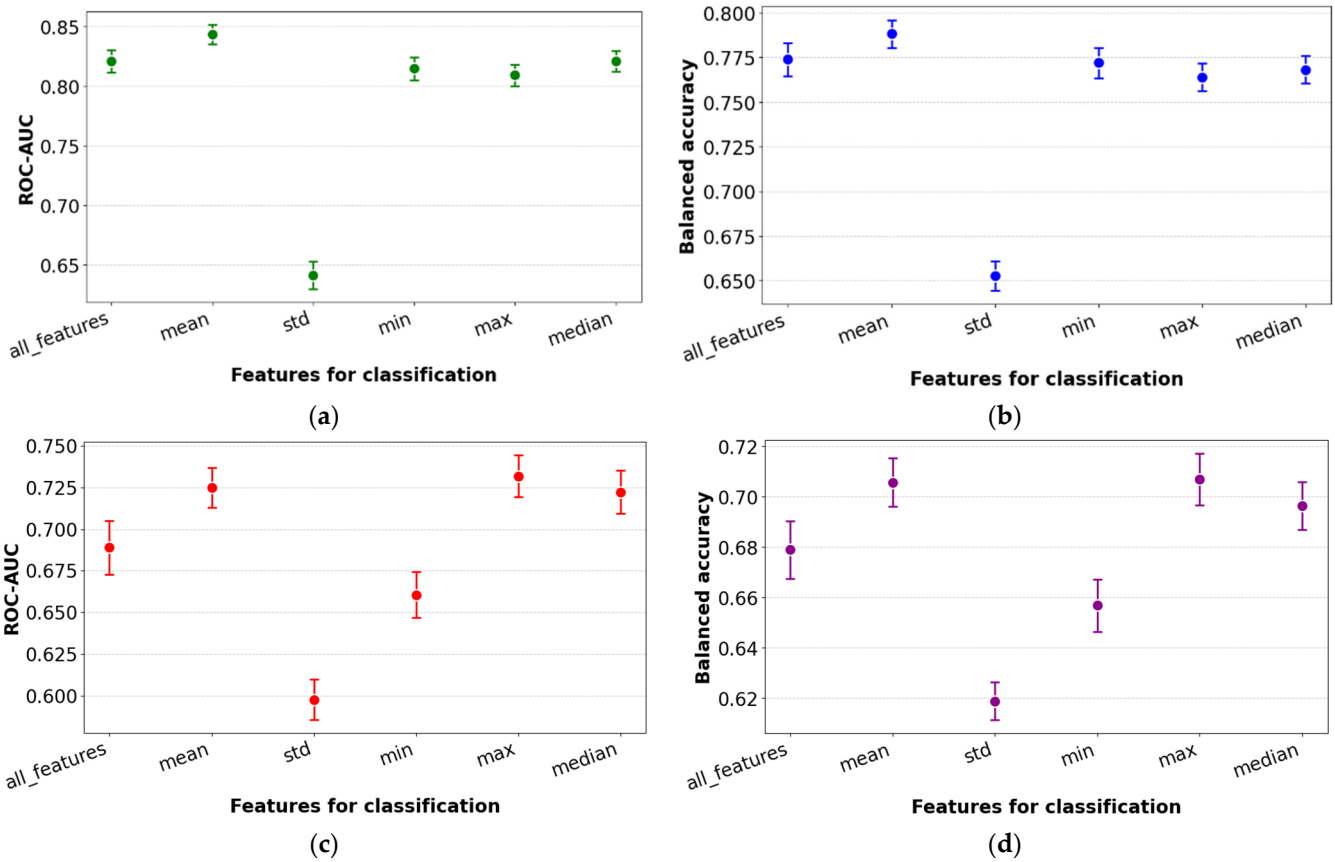


Figure 7. The classification results for canine claw samples using various statistical parameters. (**a**) ROC-AUC for Random Forest classifier; (**b**) balanced accuracy for Random Forest classifier; (**c**) ROC-AUC for Logistic Regression classifier; (**d**) balanced accuracy for Logistic Regression classifier. The error bars indicate the 95% confidence interval.

Similar to the human breast dataset, Random Forest outperformed Logistic Regression. The best metrics for Random Forest were obtained from the mean values, with an ROC-AUC of 0.844 [0.836, 0.852] and a balanced accuracy of 0.788 [0.78, 0.796]. For Logistic Regression, the best metrics are provided by the maximum.

3.2.2. Aggregation After Modeling

We applied the same aggregation strategies to the dataset of canines' nails. The results are shown in Figure 8.

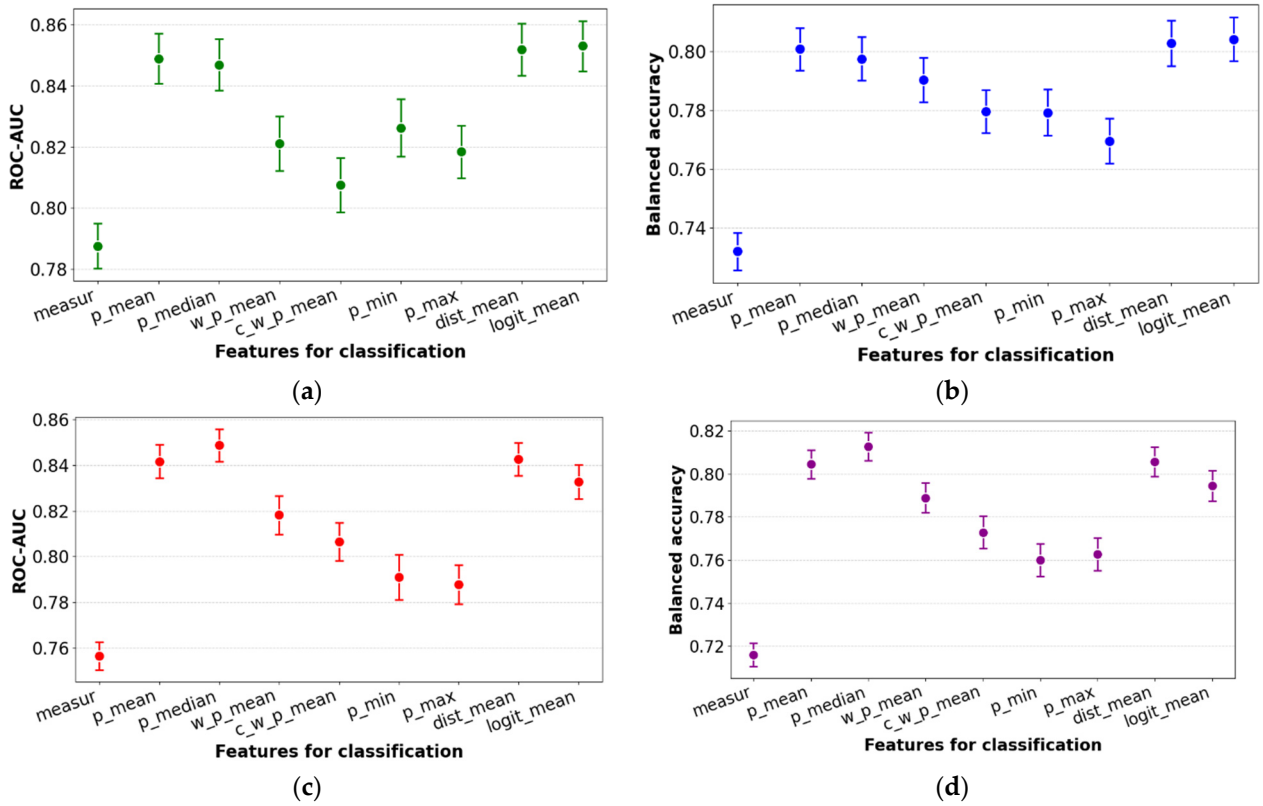


Figure 8. The classification results for canine claw samples using various aggregation procedures. (**a**) ROC-AUC for Random Forest classifier; (**b**) balanced accuracy for Random Forest classifier; (**c**) ROC-AUC for Logistic Regression classifier; (**d**) balanced accuracy for Logistic Regression classifier. The error bars indicate the 95% confidence interval.

In contrast to the human breast samples, the metrics obtained by Random Forest and Logistic Regression are similar. The best metrics for Random Forest were obtained from the logit strategy, with an ROC-AUC of 0.853 [0.845, 0.861] and a balanced accuracy of 0.804 [0.797, 0.812]. For Logistic Regression, the best metrics were provided by the median of the probabilities, with an ROC-AUC of 0.849 [0.842, 0.856] and a balanced accuracy of 0.813 [0.806, 0.819]. For both classifiers, aggregation significantly improves the metrics.

The hard prediction rule suggests that the diagnostic decision is better based on the two measurements, with both classifiers providing similar results. The metrics in this case are worse than those obtained after the aggregation, see Figure 9.

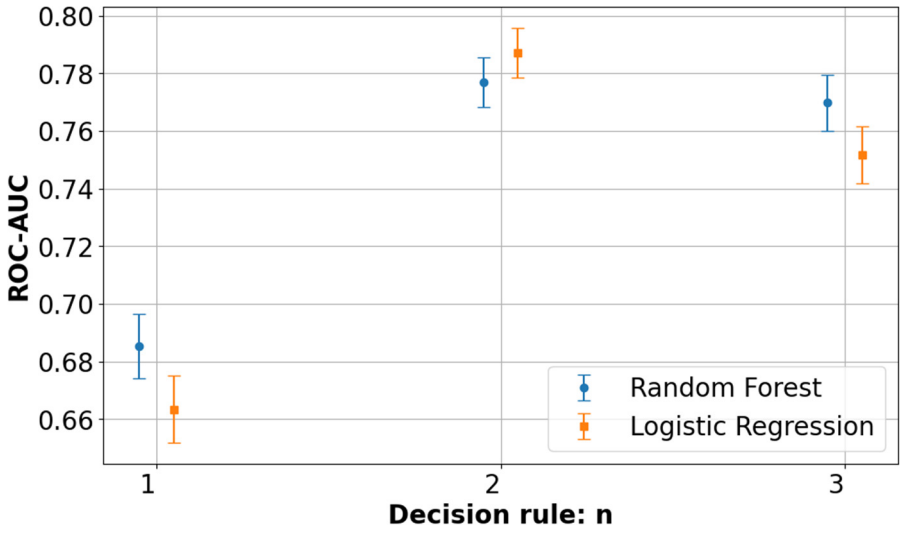


Figure 9. Hard prediction rule for canines' nails. The error bars indicate the 95% confidence interval.

4. Discussion

The goal of this work was to explore simple, straightforward aggregation methods and strategies for XRD data, applied to two datasets of XRD images, acquired from human breast biopsy samples and canine nails. We examined two main approaches: aggregating multiple measurements per patient both before and after supervised machine learning cancer/non-cancer binary classification. In both approaches and both datasets, we addressed the coefficients of the Fourier transformations of the XRD images. In the first approach, we first determined the statistical characteristics of the coefficients with respect to the measurements, and subsequently performed the classification. In the second approach, we performed classification using the Fourier coefficients of individual measurements, then aggregated the resulting cancer probabilities. We found that, in all situations, aggregation after classification yields improved metrics. For machine learning modeling, we used and compared two classifiers: Random Forest (RF) and Logistic Regression (LR). The best aggregation methods for different datasets are listed in Table 4.

Table 4. Summary of the best aggregation methods achieving maximal ROC-AUC. Aggregation approaches: 1—before modeling, 2—after modeling, and 3—hard predictions.

Data	Aggregation Approach	Best Aggregation Method (or Model)	Mean ROC-AUC	Mean BA
Biopsy WAXS	No aggregation	RF	0.840	0.774
	1	Mean + RF	0.894	0.854
	2	Logit_mean + RF	0.907	0.863
	3	n = 5, RF	0.814	
Biopsy SAXS	No aggregation	LR	0.825	0.758
	1	Not studied		
	2	C_w_p_mean + LR	0.897	0.857
	3	n = 6, LR	0.806	
Canines' nails	No aggregation	RF	0.788	0.732
	1	Mean + RF	0.844	0.788
	2	Logit_mean + RF	0.853	0.804
	3	n = 2, LR	0.787	

For the human breast samples, our examination included characteristics of collagen, triglyceride packing, lipid, and aqueous components. For the nails, we examined features of keratin molecules.

Four statistical parameters from distribution functions (mean, standard deviation, skewness, and kurtosis) were employed, and the breast study also included minimum, maximum, and median values. For the canine samples, kurtosis was not included as the number of measurements per patient was insufficient. This parameterization was implemented for both approaches (analysis pre- and post-aggregation), but the interpretations differed. In the first approach, the characteristics of the images were exploited, whereas in the second approach, the characteristics of the cancer probabilities were employed.

In the first approach, using human breast samples, aggregating and classifying WAXS data yielded better results than SAXS, and we therefore included only the WAXS metrics in this paper. Combining WAXS and SAXS data was also inferior to WAXS alone, likely due to the significant contribution from amorphous scattering, which is independent of the pathology. Means and medians for Fourier coefficients were found to produce the best Random Forest classifications (Figure 3a,b); higher-order moments were noisier, although standard deviation metrics remained reasonable.

Random Forest outperformed the Logistic Regression in all applications of the first approach. However, the best metrics for Logistic Regression are of interest because they indicate the parameters with the most well-defined clusters. For human breast samples, these characteristics were the standard deviation and minimum values. The latter corresponds to the following diagnostic rule: *if the patient's minimal coefficient belongs to the healthy cluster, then the patient is healthy*.

For the canine claw samples, Random Forest classification of the mean of the Fourier coefficients also yielded the best metrics (Figure 7a,b). The standard deviation performance is significantly poorer than for human breast tissues (compare Figure 3a,b and Figure 7a,b), indicating different cluster structures within the two datasets. For Logistic Regression, the best metrics are achieved by the maximum values. This corresponds to the following diagnostic rule: *if the patient's maximum coefficient lies within the cancerous cluster, the patient has cancer*.

In the second approach, using the human breast samples, the performance using WAXS and SAXS data was similar, and we used both sub-datasets. Random Forest was better for WAXS patterns of human breast samples (Figure 4); Logistic Regression was better for SAXS (Figure 5); and for canine nails, the two methods were similar (Figure 8). This is another indication of different cluster structures between the datasets.

For human breast samples, the best metrics were achieved using the WAXS sub-dataset, the Random Forest classifier, and the mean logit, $\ln(p_i/(1-p_i))$. Logit is well-rehearsed for the analysis of bioassays [27] and has become a commonly used statistical parameter [28]. Logit extends the probability from the [0, 1] interval to $[-\infty, \infty]$ and, as the transformation is nonlinear, the contributions from the most confident results are increased. Our use of logits was also appropriate to improve discrimination, given the probability calibration performed (although this had only a minor impact on ROC-AUC values).

Comparing the results obtained for logits with other weighting methods showed that the greater the contribution of the most confident results, the better the metric. That is, exponential weights of logit performed better than cubic or linear weights. One confirmation of the decisive contribution of the most confident predictions for diagnosis is the improved results obtained when using only the maximum probability per patient for breast cancer samples (Figures 4 and 5).

With the Logical Regression classifier, for both SAXS and WAXS, the best metrics are obtained with cubic-weighted probabilities, in which the contributions of the cancer

probabilities for individual measurements are weighted by the cube of the distance to the optimal threshold. Correspondingly, similar to the logit, the contributions of the most confident results are thus enhanced. This cubic weighting ensured that predications with higher confidence had a disproportionately larger influence. This, in turn, ensured that any diagnostic diffraction data with high confidence dominated any with low signal-to-noise ratio, thus providing selective confidence amplification.

We also performed aggregation using hard predictions, in which diagnostic decisions, not probabilities, were aggregated; i.e., we addressed the question *how many positive measurements are needed to decide that the patient has cancer?* We showed that for the human breast samples, the optimal number was 5 or 6 (out of 8–18) (Figure 6), and for canines, it was 2 (out of 3–4) (Figure 9). However, even for optimal numbers, the metrics were significantly inferior to those of the other aggregation strategies. This is consistent with the less flexible nature of hard predictions, which operate, unlike probabilities, in a binary rather than partial manner.

The post-model aggregation, in which the deep model first produced diagnostics at the most granular level (individual diffraction points), likely worked best for several reasons. Each diffraction point captured different local characteristics of the relatively heterogeneous tissues; our experimental interaction volumes were significantly smaller than the sample volumes, and different pathology states can exist within the same sample. If each measurement point were aggregated (e.g., through averaging) too early in the pipeline, then weak but diagnostically valuable signals could be washed out. Equally, scattering images with no diagnostic value (e.g., anomalous or low signal-to-noise ratio) are given little weight in the final diagnostic decision. Further, the post-model aggregation would have reduced any potential overfitting due to small patient numbers and easily accommodated the small variabilities within the numbers of diffraction data points and samples per patient.

It is currently unclear how generalizable our current results are. We consider this publication to be a formulation of the problem and an initial attempt to indicate the solutions. For now, we suggest using a similar study across all data types to identify the best aggregation method, then applying it to a diagnostic model. The methods and strategies presented in this paper are straightforward and easy to interpret. We neglected the two-step character of the measurement-to-patient transitions, omitting the sample step. This approach is justified by the need to move from a simple approach to a more complex one, identifying the most promising methods, since the number of method combinations grows as N^k, where k is the number of transitions. Therefore, the first stage of the study requires identifying the most promising aggregation methods for subsequent use in modeling multi-step transitions. Such options include at least average logits, weighted averages, and maximum and minimum probabilities per patient for post-modeling aggregation. The next step should involve using combinations of these identified methods across multiple levels of the hierarchy. In addition, more complex approaches such as stacking, ensembling, and voting should be used.

The multi-level strategies for the data's hierarchical structure will be addressed in the future. We will also consider combining various measurement types, complementing XRD with ultrasound, optical spectroscopy, blood tests, etc.

5. Conclusions

In many previous studies assessing new diagnostic assays, patient numbers are limited, whereas sample and/or assay numbers are often relatively high. This is the case described above, where the "assay" is equivalent to the X-ray scattering experiment. We examined various aggregation methods and strategies for the measurement-to-patient diagnostic

transition across two distinct datasets: human breast biopsy samples and canine nail samples. We explored two main directions, supervised machine learning classification of aggregated measurements and classification of measurements followed by aggregation, and compared two classifiers, Random Forest and Logistic Regression. We showed that aggregation improves the classification. The best metrics are achieved for both datasets when measurements are classified using Random Forest, and subsequent aggregation is performed using the logit function of the cancer probabilities.

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

Funding: The work of Alexander Alekseev is partially funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP26102549).

Institutional Review Board Statement: Tissue was collected under NRES Approval number 23/EE/0229, approved 21 November 2023, with the Keele University ethics approval granted for this research (NS-210096).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study by the Breast Cancer Now Biobank.

Data Availability Statement: The data for human breast samples [23] are available at https://zenodo.org/records/15129858 (accessed on 10 March 2026). The data for canine claw samples [26] are available at https://zenodo.org/records/14555266 (accessed on 10 March 2026).

Conflicts of Interest: Author P.L. is a shareholder of Matur UK, Ltd. and EosDx, Inc. Author A.A. is a consultant for Matur UK, Ltd. Authors K.R. and L.M. are consultants for EosDx, Inc.

References

1. Rajkomar, A.; Dean, J.; Kohane, I. Machine Learning in Medicine. *N. Engl. J. Med.* **2019**, *380*, 1347–1358. [CrossRef] [PubMed]
2. Shehab, M.; Abualigah, L.; Shambour, Q.; Abu-Hashem, M.A.; Shambour, M.K.Y.; Alsalibi, A.I.; Gandomi, A.H. Machine learning in medical applications: A review of state-of-the-art methods. *Comput. Biol. Med.* **2022**, *145*, 105458. [CrossRef]
3. An, Q.; Rahman, S.; Zhou, J.; Kang, J.J. A Comprehensive Review on Machine Learning in Healthcare Industry: Classification, Restrictions, Opportunities and Challenges. *Sensors* **2023**, *23*, 4178. [CrossRef] [PubMed]
4. Andrès, E.; Escobar, C.; Doi, K. Machine Learning and Artificial Intelligence in Clinical Medicine-Trends, Impact, and Future Directions. *J. Clin. Med.* **2025**, *14*, 8137. [CrossRef]
5. Dehbozorgi, P.; Ryabchykov, O.; Bocklitz, T.W. A comparative study of statistical, radiomics, and deep learning feature extraction techniques for medical image classification in optical and radiological modalities. *Comput. Biol. Med.* **2025**, *187*, 109768. [CrossRef]
6. Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123–140. [CrossRef]
7. Wolpert, D.H. Stacked generalization. *Neural Netw.* **1992**, *5*, 241–259. [CrossRef]
8. Mahajan, P.; Uddin, S.; Hajati, F.; Moni, M.A. Ensemble Learning for Disease Prediction: A Review. *Healthcare* **2023**, *11*, 1808. [CrossRef]
9. Müller, D.; Soto-Ray, I.; Kramer, F. An Analysis on Ensemble Learning Optimized Medical Image Classification with Deep Convolutional Neural Networks. *IEEE Access* **2022**, *10*, 66467–66480. [CrossRef]
10. Wang, Z.; Poon, J.; Sun, S.; Poon, S. Attention-based Multi-instance Neural Network for Medical Diagnosis from Incomplete and Low Quality Data. *arXiv* **2019**, arXiv:1904.04460.
11. Ilse, M.; Tomczak, J.M.; Welling, M. Attention-based Deep Multiple Instance Learning. *arXiv* **2018**, arXiv:1802.04712. [CrossRef]
12. Keskin, Z.; İnan, O.; Özberk, Ö.; Bilici, R.; Servi, S.; Çelikdelen, S.Ö.; Yıldırım, M. A Gated Attention-Based Multiple Instance Learning and Test-Time Augmentation Approach for Diagnosing Active Sacroiliitis in Sacroiliac Joint MRI Scans. *J. Clin. Med.* **2026**, *15*, 2101. [CrossRef] [PubMed]
13. Hayat, M.; Aramvith, S. Superpixel-Guided Graph-Attention Boundary GAN for Adaptive Feature Refinement in Scribble-Supervised Medical Image Segmentation. *IEEE Access* **2025**, *13*, 196654–196668. [CrossRef]

14. Mercan, C.; Aksoy, S.; Mercan, E.; Shapiro, L.G.; Weaver, D.L.; Elmore, J.G. Multi-Instance Multi-Label Learning for Multi-Class Classification of Whole Slide Breast Histopathology Images. *IEEE Trans. Med. Imag.* **2018**, *37*, 316–325. [CrossRef]
15. Tu, Y.; Lei, H.; Long, W.; Yang, Y. HAMIL: Hierarchical Aggregation-Based Multi-Instance Learning for Microscopy Image Classification. *Pattern Recognit.* **2021**, *136*, 109245.
16. Zhao, K.; Ling Yu Hung, A.; Pang, K.; Hajipour, P.; Wu, H.; Raman, S.; Sung, K. PCa-Mamba: Spatiotemporal state space models for prostate cancer detection in multi-parametric MRI. *Med. Image Anal.* **2026**, *111*, 104033. [CrossRef] [PubMed]
17. Kidane, G.; Speller, R.D.; Royle, G.J.; Hanby, A.M. X-ray scatter signatures for normal and neoplastic breast tissues. *Phys. Med. Biol.* **1999**, *44*, 1791. [CrossRef]
18. Lewis, R.A.; Rogers, K.D.; Hall, C.J.; Towns-Andrews, E.; Slawson, S.; Evans, A.; Pinder, S.E.; Ellis, I.O.; Boggis, C.R.M.; Hufton, A.P.; et al. Breast cancer diagnosis using scattered X-rays. *J. Synchrotron Radiat.* **2000**, *7*, 348–352. [CrossRef]
19. Sidhu, S.; Siu, K.K.W.; Falzon, G.; Nazaretian, S.; Hart, S.A.; Fox, J.G.; Susil, B.J.; Lewis, R.A. X-ray scattering for classifying tissue types associated with breast disease. *Med. Phys.* **2008**, *35*, 4660–4670. [CrossRef]
20. Conceicao, A.L.C.; Antoniassi, M.; Cunha, D.M.; Ribeiro-Silva, A.; Poletti, M.E. Multivariate analysis of the scattering profiles of healthy and pathological human breast tissues. *Nucl. Instrum. Methods Phys. Res. A Accel. Spectrom. Detect. Assoc. Equip.* **2011**, *652*, 870–873. [CrossRef]
21. Farquharson, M.J.; Al-Ebraheem, A.; Cornacchi, S.; Gohla, G.; Lovrics, P. The use of X-ray interaction data to differentiate malignant from normal breast tissue at surgical margins and biopsy analysis. *X-Ray Spectr.* **2013**, *42*, 349–358. [CrossRef]
22. Denisov, S.; Blinchevsky, B.; Friedman, J.; Gerbelli, B.; Ajeer, A.; Adams, L.; Greenwood, C.; Rogers, K.; Mourokh, L.; Lazarev, P. Vitacrystallography: Structural Biomarkers of Breast Cancer Obtained by X-ray Scattering. *Cancers* **2024**, *16*, 2499. [CrossRef]
23. Alekseev, A.; Shcherbakov, V.; Avdieiev, O.; Denisov, S.A.; Kubytskyi, V.; Blinchevsky, B.; Murokh, S.; Ajeer, A.; Adams, L.; Greenwood, C.; et al. Benign/Cancer Diagnostics Based on X-Ray Diffraction: Comparison of Data Analytics Approaches. *Cancers* **2025**, *17*, 1662. [CrossRef]
24. Murokh, S.; Alekseev, A.; Kubytskyi, V.; Shcherbakov, V.; Avdieiev, O.; Denisov, S.A.; Ajeer, A.; Adams, L.; Greenwood, C.; Rogers, K.; et al. X-Ray Diffraction of Collagen-Structured Water Molecules for Cancer Detection. *Molecules* **2026**, *31*, 650. [CrossRef]
25. Alekseev, A.; Yuk, D.; Lazarev, A.; Labelle, D.; Mourokh, L.; Lazarev, P. Canine Cancer Diagnostics by X-ray Diffraction of Claws. *Cancers* **2024**, *16*, 2422. [CrossRef] [PubMed]
26. Alekseev, A.; Avdieiev, O.; Murokh, S.; Yuk, D.; Lazarev, A.; Labelle, D.; Mourokh, L.; Lazarev, P. Fourier Transformation-Based Analysis of X-Ray Diffraction Pattern of Keratin for Cancer Detection. *Crystals* **2025**, *15*, 57. [CrossRef]
27. Berkson, J. Application of the Logistic Function to Bio-Assay. *J. Am. Stat. Assoc.* **1944**, *39*, 357–365.
28. Cramer, J.S. The Origins of Logistic Regression. Tinbergen Institute Working Paper No. 2002-119/4. 2002. Available online: https://www.econstor.eu/handle/10419/86100 (accessed on 30 March 2026).

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.