2023 IEEE Symposium Series on Computational Intelligence (SSCI)
Mexico City, Mexico. December 5-8, 2023

Deep learning and explainable artificial intelligence for improving specificity and detecting metabolic patterns in newborn screening

Elaine Zaunseder[1,2], Ulrike Mütze[3], Sven F. Garbade[3], Saskia Haupt[1,2], Stefan Kölker[3] and Vincent Heuveline[1,2]

Abstract— **In medical applications, artificial intelligence (AI) methods have achieved considerable progress in various areas and also in newborn screening programs. In particular, interpretable AI methods have been applied in newborn screening aiming to increase analytical specificity and predictive power of screening results. In this study, we apply ensemble and deep learning methods in newborn screening for isovaleric aciduria (IVA) on a data set containing more than 2 million newborns. We show that these methods can reduce the number of newborns falsely classified with IVA by 100% with Extreme Gradient Boosting (XGBoost), by 78.94% with Random Forest (RF), and by 78.94% with Feed Forward Neural Networks (FFNN) compared to currently applied newborn screening methods. Furthermore, we show how explainable AI (XAI) methods can be used to interpret these black-box classification results and further apply them for potential biomarker discovery. The XAI methods reveal that besides the biomarker isovaleryl carnitine (C5), the birth year and the amino acid tryptophan (Trp) are influential in reducing the false positive rate. By this, we show that ensemble and deep learning could be highly beneficial in newborn screening and could have a major impact on newborns and their families, as it reduces false positive screening results and guides new directions for future research in this field.**

I. Introduction

In the medical domain, ensemble and deep learning methods have been successfully applied in various areas [1], [2] and especially in disease prediction tasks such as diabetes prediction [3] and prediction of colorectal cancer among patients [4]. Newborn screening programs act worldwide to identify pre-symptomatic newborns suffering from severe rare metabolic diseases by analyzing different metabolite concentrations in the newborns' blood samples [5]. Due to the low prevalence of screened diseases, these results have to be highly accurate, aiming at high sensitivity and specificity, to reliably identify all newborns with a disease and reduce the number of false positives. Identification by newborn screening allows early, ideally presymptomatic start of treatment of affected newborns. Isovaleric aciduria (IVA) is an organic aciduria that is included in newborn screening disease panels. In its severest form, affected individuals present with life-threatening (neonatal) metabolic compensations [6]. In 2005, IVA became a target disease in German regular newborn screening which enabled earlier specialized treatment and, thus, reduced neonatal mortality of affected individuals [7]. However, the newborn screening for IVA is hampered by an increasing number of false positives due to the increasing use of pivmecillinam, an antibiotic used in urinary tract infections in pregnant women [8].

Recently, different machine learning (ML) methods have been applied to newborn screening data to improve the classification accuracy by reducing false positive rates and identifying so far unknown metabolic patterns by relying on complex feature combinations instead of predefined single cut-off values [9]. However, deep learning methods such as neural networks (NN) and extreme gradient boosting (XGBoost) which are often the best-performing methods for classification, have only rarely been used [10]. Their high complexity and corresponding lack of interpretability often prevent their usage in newborn screening. Hence, new computational approaches to improve newborn screening and interpret black-box deep learning models are needed to reduce the newborns' and families' burden of false positives including over-treatment, and gain insights into biomarker patterns.

II. Related Work

Several ML methods have been applied in newborn screening, as a recent systematic literature review shows [10]. Among the previously applied ML methods, logistic regression (LR) and support vector machine showed good performance for newborn screening classification in single and comparative studies [10]. Specially for the newborn screening conditions phenylketonuria [11], [12], [13], methylmalonic aciduria [11], [14], [9] and medium-chain acyl-CoA dehydrogenase deficiency [11], [15], [16] various studies showed improved classification results. A recent study on ML for IVA developed a new method based on LR and linear discriminant analysis (LDA) which reduced the false positive rate by nearly 70% and did not investigate the application of ensemble or deep learning techniques for IVA [17]. Until now, Feed Forward Neural Networks (FFNN) [18], [11], boosting methods [14], and Random Forest (RF) [19], [9] are only used in some comparative studies for other newborn screening conditions. In addition, pattern recognition techniques are applied in newborn screening to detect patterns within the metabolite concentrations. For instance, built-in decision functions [18], [20], [15], [19] and discriminatory thresholds [18], [11] were used to identify

[1]Engineering Mathematics and Computing Lab, Heidelberg University, 69120 Heidelberg, Germany <firstname.lastname>@uni-heidelberg.de

[2] Data Mining and Uncertainty Quantification (DMQ), Heidelberg Institute for Theoretical Studies (HITS), 69118 Heidelberg, Germany

[3] Division of Child Neurology and Metabolic Medicine, Center for Child and Adolescent Medicine, Heidelberg University Hospital, 69120 Heidelberg, Germany <firstname.lastname>@med.uni-heidelberg.de

978-0-7381-4408-5/23/$31.00 ©2023 IEEE

important biomarkers from the ML classification. For non-interpretable ML methods, model agnostic approaches such as mean decrease accuracy were applied to identify the contribution of individual biomarkers to the overall classification result [21], [9]. Recently, eXplainable Artificial Intelligence (XAI) techniques have been applied to provide human-understandable interpretations of deep learning algorithms in the medical domain [22]. For instance, SHapley Additive exPlanations (SHAP) was used for melanoma detection in skin imaging [23], or the SurvLIME was developed for explaining machine learning survival models [24]. However, also in newborn screening, XAI techniques seem to be a promising approach as they could enable a better understanding of the underlying biochemical mechanisms and support the application of deep learning techniques for disease diagnosis.

Our Contribution Here, we show the application of ensemble and deep learning methods to help to improve newborn screening for IVA and how XAI techniques can be utilized to improve interpretability for clinicians and reveal influential biomarkers to guide future research directions in this field.

III. Material and Methods

This section presents the medical data sets, experimental setup, ensemble, and deep learning methods, and details on applying XAI methods for interpretability and pattern recognition.

A. Data sets

The applied anonymized data set in this study was retrieved from the NBS laboratory at Heidelberg University Hospital (UKHD), where about 20% of all newborns born in Germany are screened [25]. The UKHD data protection officer checked that the anonymized set of NBS variables is in accordance with the European General data protection regulation (GDPR). The whole data set contained NBS profiles of 2,237,142 newborns born between 2002 and 2021 containing 53 numerical and categorical features. These included 48 metabolite concentrations and five additional variables such as sex, birth weight, age at blood sample, age at sample arrival, and gestational age for each newborn. The confirmed diagnosis was set as the target variable. The extracted data set was cleaned by removing missing values and not interpretable entries, to ensure high data quality. After consultation with clinical experts, the following ranges were defined to exclude NBS profiles with implausible values: Birth weight: $1000-6000$ g; gestational age: $32-42$ weeks, age at sampling: $36-120$ hours, age at sample arrival: $0-20$ days and metabolite concentrations: $0-50{,}000\mu$mol/l [17]. These data cleaning steps resulted in a highly imbalanced data set, which we name "full data set" in this study, with 2,106,090 NBS profiles, including only 28 confirmed IVA cases. Additionally, we extracted a subset of the full data set, which we name "diagnosis subset" containing the 28 confirmed IVA cases and all 103 healthy newborns with initial IVA positive screening results, i.e., identified as false positives. Hence, an ML classification on the diagnosis subset can simulate the scenario, where the algorithm is only applied after the traditional newborn screening to reduce false positive screening results further. In order to be able to compare our results to the traditional newborn screening we do not apply sampling algorithms [11], [14] or reduced data sets to overcome data imbalance [15] since these change the sick-to-control ratio within the data sets.

B. Experimental setup

The experimental setup describes how algorithms were developed and optimized. All experiments were performed on an Nvidia GeForce RTX 3090 Ti with the deep learning platform Torch, deep learning backend Torch-gpu, Cuda version 12.0, and CuDNN version 11.7. as well as the Python library scikit-learn [26]. Both data sets are subject to data imbalance, where the true positives are in the minority. To overcome this data imbalance, we used a grid search on the model parameters of the methods to find the optimal hyperparameters. The minority class weight parameter w_1 can be used to penalize a miss-classification of a true positive in the cross entropy loss function. The data sets are randomly split 65-15-20 into training, validation, and test set, using a stratified splitting method to ensure an appropriate proportion of IVA samples in each data set. The models are then evaluated on the number of false positives (FP) and false negatives (FN) in each data set as well as the mean sensitivity S_n and specificity S_p from ten times repeated 5-fold cross-validation (CV),

$$S_n = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \quad , \quad \text{and} \quad S_p = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \quad ,$$

with true negatives (TN), and true positives (TP).

C. Ensemble and deep learning models

The sub-field of ensemble and deep learning models describes complex algorithms that learn from large amounts of data and can make well-founded decisions. In this study, we focus on FFNN, XGBoost, and RF which achieved good results for classification on tabular data [27]. FFNNs try to mimic the signaling processes in the human brain and exclusively pass information forward through multilayer perceptrons and were applied in this study, as the newborn screening profiles are independent of each other [18], [11]. The grid search for hyperparameter optimization of the FFNNs resulted in two different architectures for the two data sets, Table I. It shows that especially the class weight parameters differ between the two architectures, as the data

TABLE I
Overview of hyperparameters for FFNN architectures for diagnosis subset and full data set with four layers each.

	Diagnosis subset	Full data
Class weight w_0	0.64	0.520833
Class weight w_1	2.34	37608.75
Neurons	[100,53,6,2]	[256,128,64,2]
Optimizer	SGD	RSMEprop

imbalance is much larger in the full data set than in the diagnosis subset. RF is an ensemble method combining several randomly initialized decision trees to one powerful classifier [9]. A grid search determined 50 estimators, a maximal tree depth of 4 and class weights $w_0 = 1$ and $w_1 = 10$ as optimal hyperparameters for the diagnosis subset. Whereas for the full data set, 61 estimators, a maximal tree depth of 3, as well as $w_0 = 1$ and $w_1 = 50{,}000$, were optimal. XGBoost is a gradient-boosting library for reducing bias and variance where each weak learner tries to correct the model predictions of its predecessors [28]. We used the gbtree booster and optimized the numbers of estimators, tree depth, and scale positive weight to account for the imbalanced data. For both data sets, 16 estimators and a tree depth of 3 features were optimal, and a scaled weight $w_s = 1$ for the diagnosis subset and $w_s = 6$ for the full data set were the best hyperparameters.

D. XAI methods for interpretability

Highly complex ML methods such as ensemble learners or NNs are not inherently interpretable, which leads to the unavoidable trade-off between the accuracy and interpretability of a model's output. Hence, simpler explanation models are desired to achieve accurate results and an intuitive understanding of the complex classifiers output [29]. We used the SHAP values [29] to quantify the interactions between features that lead to a prediction by fairly distributing the payout among the features. These values can then be interpreted as which features were most influential for a certain classification outcome.

E. Pattern recognition with LIME

Another commonly applied XAI method is the Local Interpretable Model-agnostic Explanations (LIME) that explains the predictions of any classifier by learning an interpretable model locally around the prediction [30]. It can be applied to tabular data, such as newborn screening data, where the prediction explanation is given as importance scores for each feature. By this, we can gain insights into the model's decision-making process for classifying one newborn. We employed LIME using a novel configuration to interpret the classification model and gain insights into the underlying metabolic patterns evident in the concentrations of metabolites. Specifically, we applied a cumulative approach using LIME to analyze the group of true positive predictions, as identifying similarities among these individuals enables us to gain insights into metabolic patterns within newborns with confirmed IVA. For every newborn i from a group of n newborns, we obtained a LIME score l_{ji} for every feature j. We analyzed this with two cumulative LIME scores, which add the positive L_j^+ and negative contributions L_j^- separately for every feature j,

$$L_j^+ = \sum_i l_{ij}, \qquad i = \{i = 1,...,n | l_{ij} > 0\},$$
$$L_j^- = \sum_i l_{ij}, \qquad i = \{i = 1,...,n | l_{ij} < 0\}.$$

Here, a high L_j^+ score indicates that feature j makes a positive disease prediction more likely, whereas a low L_j^- indicates the opposite.

IV. Results

In this section, we present the results of the ensemble and deep learning classification for newborn screening for IVA, as well as the XAI results and the patterns identified by the LIME-based pattern recognition scheme.

A. Ensemble and deep learning classification

Based on the presented experimental setup, we trained and optimized the classification methods to improve newborn screening for IVA on both data sets and show the classification results in Table II. On the diagnosis subset, the FFNN, RF, and XGBoost classification enable a huge reduction of false positives on the training, validation, and test set compared to 103 false positives in traditional newborn screening. XGBoost reduces the false positive rate of the test set by 100%, FFNN by 89.47%, and RF by 78.94%, Table II. All three methods wrongly classify at least one newborn with IVA as insusceptible, whereas the traditional newborn screening has a sensitivity of 100%. The cross-validation results for the ensemble learners show high mean sensitivity 94.49% for XGB and 92.89% for RF as well as high mean specificity 97.38% for XGB and 92.65% for RF. However, the FFNN shows less reliable results in cross-validation, decreasing the mean sensitivity to 43.36% and specificity to 71.24%, Table II.

On the full data set containing more than 2 million newborns, all three methods decrease the number of false positives, XGBoost decreasing the false positive rate of the test set by 94.44%, FFNN by 77.78%, and RF by 100%. This results in an increase of the specificity from 99.995% in traditional newborn screening to 99.999% with XGBoost, FFNN, and RF classification, Table II. However, on the full data set, all methods show an increased number of false negative predictions, Table II. Similar to the results on the diagnosis subset, the ensemble methods demonstrate higher mean sensitivity, 93.73% for XGB and 79.46% for RF compared to the FFNN, 65.36%. Overall, the ensemble methods XGBoost and RF obtain good classification results more reliably than FFNN, and on both data sets, XGBoost

TABLE II
CLASSIFICATION RESULTS ON TRAINING, VALIDATION, AND TEST SET.

Method	Train		Validation		CV		Test	
	FN	*FP*	*FN*	*FP*	$S_n(\%)$	$S_p(\%)$	*FN*	*FP*
DIAGNOSIS SUBSET								
Traditional	0	72	0	12	100	0	0	19
XGBoost	0	0	1	3	94.491	97.382	1	0
FFNN	0	0	1	3	43.357	71.237	1	2
RF	0	3	2	0	92.885	92.648	1	4
FULL DATA SET								
Traditional	0	67	0	18	100	99.995	0	18
XGBoost	0	0	1	1	93.727	99.999	2	1
FFNN	2	7	1	0	65.357	99.999	2	4
RF	0	3	4	1	79.455	99.999	4	0

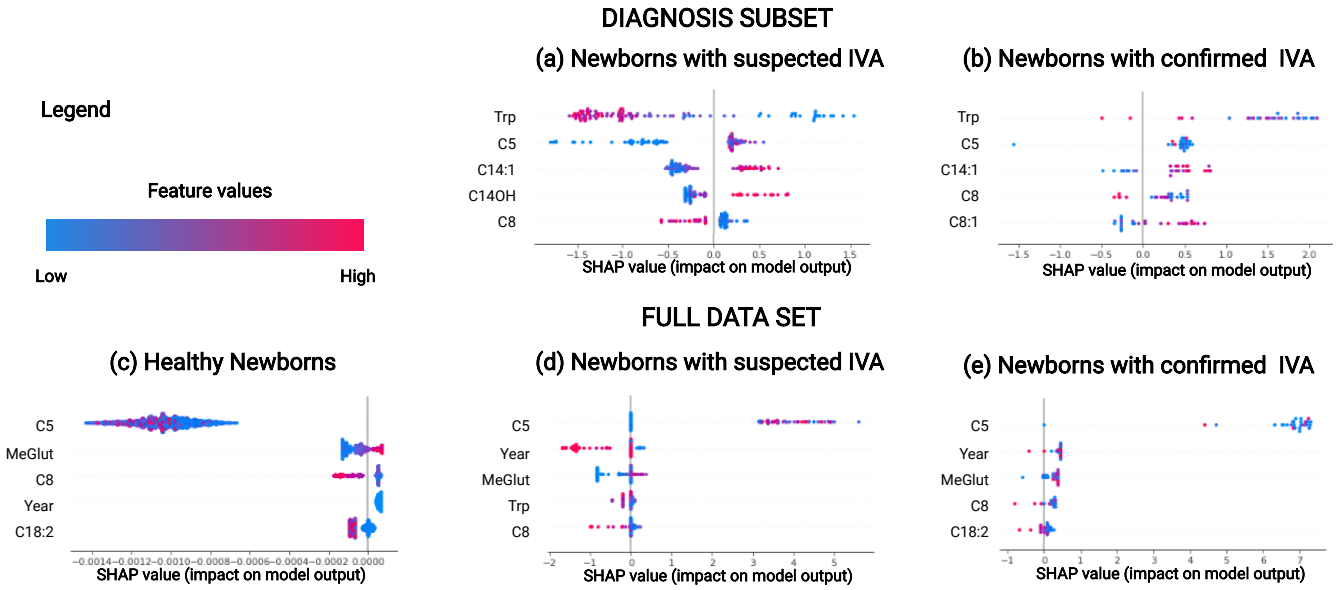

Fig. 1. SHAP values for the diagnosis subset (a), (b) and the full data set (c)-(e) for the groups of healthy newborns, newborns with suspected IVA, and newborns with confirmed IVA presenting the five highest features. The data points are color-coded depending on the feature value, where blue data points are low feature values, and red data points are high feature values. **Abbreviations:** C14:1 - tetradecenoylcarnitine, C14OH - 3-OH-tetradecanoylcarnitine, C18:2 - octadecadienoylcarnitin, C8 - octanoylcarnitine, C8:1 - octenoylcarnitine, C5 - isovalerylcarnitine, MeGlut - 3-methylglutarylcarnitine, Trp - tryptophan, Year - birth year.

improves the mean sensitivity and specificity more than RF, therefore XGBoost's decision-making will be examined in the following sections.

B. Interpretation of XGBoost classification with SHAP

As newborn screening is a highly sensitive, population-based preventive program, the interpretation and explainability of ensemble and deep learning classification procedures are important to optimize the specificity and predictive power of screening results. We applied a SHAP tree explainer method on XGBoost trained on the two data sets, full data and diagnosis subset. We divided the data sets into the groups of healthy newborns, newborns with suspected IVA, and newborns with confirmed IVA to gain insights into the differences and similarities of these groups, Fig. 1. Traditional newborn screening only considers the primary marker C5 for IVA, whereas the ensemble model considers all metabolite concentrations. For the diagnosis subset, feature C5 only has the second highest influence, whereas the impact of tryptophan (Trp) is the most important. In Fig. 1 (a), a lower amount of Trp coincides with a positive impact on the model output, meaning a higher likeliness to be classified as newborn with IVA and a higher amount of Trp relates to a negative impact on the model output and, hence, more likely to be classified as a healthy newborn. Furthermore, low feature values of C5 impact the model's output in a negative direction and make a healthy classification more likely, which coincides with a high value of C5 being a primary marker for IVA. Fig. 1 (c) shows that the XGBoost classifier on the full data set accounts for C5 being the primary marker for IVA as the value of this group strongly influences the model to classify the newborn as normal. Whereas for the group of suspected positive newborns and newborns with confirmed IVA, we see a pattern for newborns with higher C5 values influencing a positive disease prediction. Furthermore, the feature birth year is strongly negatively correlated with higher SHAP values, which coincides with the birth year being highly correlated with the increasing use of pivmecillinam since 2016 [8], leading to an increase in false-positive screening results. Interestingly, Trp is one of the top influential features for newborns with suspected and confirmed IVA but not for healthy newborns. Moreover, M3-methylglutarylcarnitine (MeGlut) is influential for all three groups of the full data set.

C. Pattern recognition with LIME

Besides trying to interpret the ML models using XAI, we aim at gaining insight into underlying metabolic patterns. In Fig. 2 (a), the LIME score for an individual newborn with IVA, which was correctly classified by the XGB model trained on the full data set, is shown. For this patient, C5 and birth year are the most influential features for the correct prediction of IVA. To further understand the classifications and interpret the XGBoost prediction, we provide the cumulative LIME score for interpreting the data subgroups. Hence, we investigate the five metabolites j with the most significant difference between negative and positive cumulative LIME scores L_j^- and L_j^+ for all newborns with IVA that XGBoost correctly classified, Fig. 2 (b), (c). Similar to the SHAP evaluation in the diagnosis subset, the metabolite Trp shows a high cumulative positive LIME score of 18.39 and no cumulative negative LIME score, which marks this feature

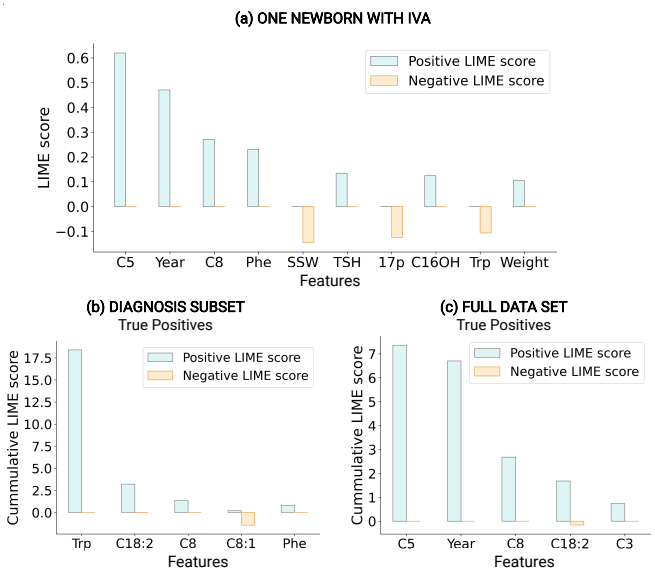

Fig. 2. LIME scores for an individual newborn (a) for ten highest features and cumulative LIME scores for a group of correctly classified newborns with IVA in the diagnosis subset (b) and the full data set (c) divided into positive and negative LIME scores for five highest features. **Abbreviations:** 17p - 17-hydroxyprogesterone, C14:1 - tetradecenoylcarnitine, C14OH - 3-OH-tetradecanoylcarnitine, C16OH - 3-OH-hexadecanoylcarnitine, C18:2 - octadecadienoylcarnitin, C3 - propionylcarnitine, C5 - isovalerylcarnitine, C8 - octanoylcarnitine, C8:1 - octenoylcarnitine, MeGlut - 3-methylglutarylcarnitine, Phe - phenylalanine, SSW - gestational age, Trp - tryptophan, TSH - thyroid-stimulating hormone, Year - birth year.

as highly influential for the correct classification as newborn with confirmed IVA, Fig. 2 (b). All other metabolites show far lower cumulative LIME scores. For the full data set, both the metabolite C5 (7.36) and the feature birth year (6.7) show high cumulative positive LIME scores and zero negative LIME score, while all other features have smaller values, Fig. 2 (c).

V. Summary and Discussion

Newborn screening programs are essential for identifying newborns with treatable rare diseases worldwide. In this study, we show that ensemble and deep learning methods can be applied to improve the specificity of newborn screening for IVA, which is hampered in Germany by an increasing number of false positive screening results [8]. In particular, on the diagnosis subset, which simulates applying ML as an additional step after the traditional newborn screening, the number of newborns falsely classified as newborns with IVA on the test set can be reduced by 100% with XGBoost, 89.47% with FFNN and 78.94% with RF, Table II. However, when evaluating the cross-validation results, XGBoost proved to be the more robust method obtaining $S_n = 93.727\%$ sensitivity and $S_p = 99.999\%$ specificity on the full data set compared to a sensitivity of 100% and specificity of 99.995% with traditional newborn screening and sensitivity of 100% and specificity of 99.998% with LR [17]. As these methods are not inherently interpretable, we apply XAI techniques SHAP and LIME on XGBoost to interpret the results and identify underlying biochemical patterns. The evaluation of the SHAP values confirmed biological knowledge, such as highlighting C5, the known primary marker of standard newborn screening for IVA, as a significant feature for the full data set [6], [7]. Also, birth year was identified as an influential feature, which is explained by the increasing number of false positives in recent years due to the use of pivmecillinam as an antibiotic in pregnant women since its authorization in Germany in 2016 [8]. Furthermore, the cumulative LIME evaluation identified the amino acid Trp as an influencing feature for the diagnosis subset, especially for the correct classification of former false positives and hence, for improving the specificity. However, a biochemical explanation for this revelation is difficult and should be investigated by clinical experts and evaluated on further IVA data sets in future studies. These results could be a direction for further medical research on IVA.

Besides improving specificity, sensitivity is an important measure in newborn screening. However, although XGBoost achieved high sensitivity values, it did not obtain 100% sensitivity in cross-validation such as traditional screening and classical ML [17], which could be due to the low number of IVA cases in the data set. Hence, more IVA data from different newborn screening centers should be included in future studies to ensure higher positive sample sizes and validate the identified patterns. Furthermore, newborn screening for other rare diseases could also benefit from an application of ML to improve specificity, as data is collected for decades for most conditions. Overall, the reduction of false-positive newborn screening results has several positive effects as human and material resources could be reduced, hard- and software costs are low, and newborn screening laboratories would have less effort to reporting and tracking the confirmatory diagnostics. Therefore, this approach should be evaluated in daily practice in parallel to traditional newborn screening to assess the possible cost-effectiveness. Besides the performance, the interpretability of ML methods becomes an important topic, as it is unclear how black-box methods can be applied in the clinical context and how these methods may be controlled and accepted by patients and society [31]. A more frequent application of XAI methods for newborn screening could enhance the understanding and interpretability of ensemble and deep learning methods, leading to a higher acceptance of these.

In summary, our study demonstrates that the utilization of ensemble and deep learning techniques leads to increased specificity in IVA classification within the presented framework. Moreover, the outcomes obtained through XAI provide promising insights into result interpretation, offering novel prospects for future investigations in newborn screening research.

Acknowledgment

This research was funded by the Klaus Tschira Foundation through the Informatics for Life project and the Dietmar Hopp Foundation, St. Leon Rot, Germany (grant numbers

2311221, 1DH2011117 and 1DH1911376). The authors confirm independence from the sponsors; the content of the article has not been influenced by the sponsors.

REFERENCES

[1] S. McKinney, M. Sieniek, and V. Godbole, "International evaluation of an ai system for breast cancer screening," *Nature*, vol. 577, p. 89–94, 2020.

[2] D. Ardila, A. Kiraly, and S. Bharadwaj, "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography," *Nature Medicine*, vol. 25, p. 954–961, 2019.

[3] M. Li, X. Fu, and D. Li, "Diabetes prediction based on XGBoost algorithm," *IOP Conference Series: Materials Science and Engineering*, vol. 768, no. 7, 2020.

[4] R. Amirkhan, M. Hoogendoorn, M. E. Numans, and L. Moons, "Using recurrent neural networks to predict colorectal cancer among patients," in *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2017, pp. 1–8.

[5] B. Therrell, C. Padilla, J. Loeber, I. Kneisser, A. Saadallah, G. Borrajo, and J. Adams, "Current status of newborn screening worldwide: 2015," *Seminars in Perinatology*, vol. 39, no. 3, pp. 171–187, 2015.

[6] R. Ensenauer, J. Vockley, J. Willard, J. C. Huey, J. O. Sass, S. D. Edland, B. K. Burton, S. A. Berry, R. Santer, S. Grünert, H.-G. Koch, I. Marquardt, P. Rinaldo, S. Hahn, and D. Matern, "A common mutation is associated with a mild, potentially asymptomatic phenotype in patients with isovaleric acidemia diagnosed by newborn screening," *The American Journal of Human Genetics*, vol. 75, no. 6, pp. 1136–1142, 2004.

[7] U. Mütze, L. Henze, F. Gleich, M. Lindner, S. C. Grünert, U. Spiekerkoetter, R. Santer, H. Blessing, E. Thimm, R. Ensenauer, J. Weigel, S. Beblo, M. Arélin, J. B. Hennermann, T. Marquardt, I. Marquardt, P. Freisinger, J. Krämer, A. Dieckmann, N. Weinhold, M. Keller, M. Walter, K. A. Schiergens, E. M. Maier, G. F. Hoffmann, S. F. Garbade, and S. Kölker, "Newborn screening and disease variants predict neurological outcome in isovaleric aciduria," *Journal of Inherited Metabolic Disease*, vol. 44, no. 4, pp. 857–870, 2021.

[8] S. Murko, A. D. Aseman, F. Reinhardt, G. Gramer, J. G. Okun, U. Mütze, and R. Santer, "Neonatal screening for isovaleric aciduria: Reducing the increasingly high false-positive rate in germany," *JIMD Reports*, vol. 64, no. 1, pp. 114–120, 2023.

[9] G. Peng, Y. Tang, T. Cowan, G. Enns, H. Zhao, and C. Scharfe, "Reducing false-positive results in newborn screening using machine learning," *Int J Neonatal Screen*, vol. 6, no. 1, 2020.

[10] E. Zaunseder, S. Haupt, U. Mütze, S. Garbade, S. Kölker, and V. Heuveline, "Opportunities and challenges in machine learning-based newborn screening—a systematic literature review," *JIMD Reports*, 2022.

[11] C. Baumgartner and D. Baumgartner, "Biomarker discovery, disease classification, and similarity query processing on high-throughput ms/ms data of inborn errors of metabolism," *Journal of biomolecular screening*, vol. 11, no. 1, pp. 90–99, 2006.

[12] W. Chen, H. Chen, Y. Tseng, K. Hsu, S. Hsieh, Y. Chien, W. Hwu, and F. Lai, "Newborn screening for phenylketonuria: Machine learning vs clinicians," *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 798–803, 2012.

[13] Z. Zhu, J. Gu, G. Genchev, X. Cai, Y. Wang, J. Guo, G. Tian, and H. Lu, "Improving the Diagnosis of Phenylketonuria by Using a Machine Learning–Based Screening Model of Neonatal MRM Data," *Frontiers in Molecular Biosciences*, vol. 7, 2020.

[14] B. Lin, J. Yin, Q. Shu, S. Deng, Y. Li, P. Jiang, R. Yang, and C. Pu, "Integration of machine learning techniques as auxiliary diagnosis of inherited metabolic disorders: Promising experience with newborn screening data," *Collaborative computing: Networking, applications and worksharing*, vol. 292, pp. 334–349, 2019.

[15] S. Ho, Z. Lukacs, G. Hoffmann, M. Lindner, and T. Wetter, "Feature Construction Can Improve Diagnostic Criteria for High-Dimensional Metabolic Data in Newborn Screening for Medium-Chain Acyl-CoA Dehydrogenase Deficiency," *Clinical Chemistry*, vol. 53, no. 7, pp. 1330–1337, 2007.

[16] T. Van den Bulcke, P. Vanden Broucke, V. Van Hoof, K. Wouters, S. V. Broucke, G. Smits, E. Smits, S. Proesmans, T. V. Genechten, and F. Eyskens, "Data mining methods for classification of medium-chain acyl-coa dehydrogenase deficiency (mcadd) using non-derivatized tandem ms neonatal screening data," *J. of Biomedical Informatics*, vol. 44, no. 2, p. 319–325, 2011.

[17] E. Zaunseder, U. Mütze, S. F. Garbade, S. Haupt, P. Feyh, G. F. Hoffmann, V. Heuveline, and S. Kölker, "Machine learning methods improve specificity in newborn screening for isovaleric aciduria," *Metabolites*, vol. 13, no. 2, 2023.

[18] C. Baumgartner, C. Böhm, D. Baumgartner, G. Marini, K. Weinberger, B. Olgemöller, B. Liebl, and A. A. Roscher, "Supervised machine learning techniques for the classification of metabolic disorders in newborns," *Bioinformatics*, vol. 20, no. 17, pp. 2985–2996, 2004.

[19] S. Zarin Mousavi, M. Mohammadi Zanjireh, and M. Oghbaie, "Applying computational classification methods to diagnose congenital hypothyroidism: A comparative study," *Informatics in Medicine Unlocked*, vol. 18, p. 100281, 2020.

[20] C. Baumgartner, C. Böhm, and D. Baumgartner, "Modelling of classification rules on metabolic patterns including machine learning and expert knowledge," *Journal of Biomedical Informatics*, vol. 38, no. 2, pp. 89–98, 2005.

[21] G. Peng, P. Shen, N. Gandotra, A. Le, E. Fung, L. Jelliffe-Pawlowski, R. W. Davis, G. M. Enns, H. Zhao, T. M. Cowan, and C. Scharfe, "Combining newborn metabolic and DNA analysis for second-tier testing of methylmalonic acidemia," *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, vol. 21, no. 4, pp. 896–903, 2019.

[22] A. Singh, S. Sengupta, and V. Lakshminarayanan, "Explainable deep learning models in medical image analysis," *Journal of Imaging*, vol. 6, no. 6, 2020.

[23] K. Young, G. Booth, B. Simpson, R. Dutton, and S. Shrapnel, "Deep neural network or dermatologist?" in *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*. Springer International Publishing, 2019, pp. 48–55.

[24] M. S. Kovalev, L. V. Utkin, and E. M. Kasimov, "Survlime: A method for explaining machine learning survival models," *Knowledge-Based Systems*, vol. 203, p. 106164, 2020.

[25] U. Mütze, S. Garbade, G. Gramer, M. Lindner, P. Freisinger, S. C. Grünert, J. Hennermann, R. Ensenauer, E. Thimm, J. Zirnbauer, M. Leichsenring, F. Gleich, F. Hörster, K. Grohmann-Held, N. Boy, J. Fang-Hoffmann, P. Burgard, M. Walter, G. F. Hoffmann, and S. Kölker, "Long-term Outcomes of Individuals With Metabolic Diseases Identified Through Newborn Screening," *Pediatrics*, vol. 146, no. 5, 2020.

[26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[27] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," *Information Fusion*, vol. 81, pp. 84–90, 2022.

[28] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: Association for Computing Machinery, 2016, p. 785–794.

[29] S. Lundberg and S. Lee, "A unified approach to interpreting model predictions," *Proceedings of the 31st International Conference on Neural Information Processing Systems*, p. 4768–4777, 2017.

[30] M. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 1135–1144, 2016.

[31] M. Arnold, "Teasing out artificial intelligence in medicine: An ethical critique of artificial intelligence and machine learning in medicine," *Journal of bioethical inquiry*, vol. 18, no. 1, pp. 121–139, March 2021.