A Framework for Diagnostic Decision Support in Rheumatic and Autoimmune Diseases Using Machine Learning

Akeel Ibrahem Bdrany[1], Mohammed Layth Talal[2], Fatima Ismael Bdrany[1], Qahtan Majeed Yas[3] and Marwan Badran Mohammed[4]

[1]*Department of Computer Science, College of Agriculture, University of Diyala, 32001 Baqubah, Iraq*
[2]*Department of Computer Science, College of Administration and Economics, University of Diyala, 32001 Baqubah, Iraq*
[3]*Department of Artificial Intelligence, College of Engineering for Artificial Intelligence Technology, University of Diyala, 32001 Baqubah, Iraq*
[4]*Department of Computer Science, College of Science, Al-Nahrain University, 10070 Baghdad, Iraq*
{akeelbdrany, mohammedeco, fatimabdrany, qahtan.myas}@uodiyala.edu.iq, marwan.badran@nahrainuniv.edu.iq

Keywords: Rheumatology, Autoimmune Disease, Diagnostic Decision Support, Machine Learning, XGBoost, CatBoost, Hierarchical Classification, Differential Diagnosis, Calibration.

Abstract: This study presents a machine-learning framework that has been created using routine lab data as a tool to support diagnostic decisions for rheumatologic and autoimmune diseases. A publicly available Iraqi dataset was used, comprising 12,085 biologically reliable patients, 14 diagnostic features, and 7 diagnostic examples. Both the XGBoost and CatBoost primary classifiers were employed, along with a hierarchical integration of multiple expert predictive models. The results were confirmed when the model was tested on all patient samples not used in training. The proposed machine learning framework achieved an accuracy of 0.8606; an overall F1 mean of 0.8535; a balanced accuracy of 0.8452; a Matthews's correlation coefficient (MCC) of 0.8339; and a two-highest accuracy of 0.9880. These results support differential diagnosis in cases where specific tests may not be available. The implication of the study, based on a machine learning framework, proved to be a valuable tool for supporting diagnostic decision-making, as it provided an improved classification of probabilities, but it does not replace the physician's final decision in diagnosis.

1 INTRODUCTION

Rheumatic and autoimmune diseases comprise a heterogeneous group of inflammatory disorders that overlap clinically and laboratorial at many stages of diagnosis, especially when routine markers such as ESR, CRP, and selected autoantibodies are used. Differential diagnosis becomes even more difficult when serological profiles converge across diseases or when cases are seronegative, which may delay correct referral or timely initiation of appropriate treatment [1]-[3]. The Data in Brief published Iraqi dataset is specifically designed to enable testing of machine-learning-based clinical decision support models. The dataset contains 12085 de-identified patient records and 14 different features that fall into 7 Diagnostic Classes [1]. Subsequent research based on the same dataset has suggested that Tree-Based Ensemble Methods produce relatively good results. For example, Mahdi et al. found that using Gradient Boosting combined with FDOSM and LIME produced an accuracy of 86.89%. However, Elmesiry et al. found that the best accuracy achieved on their test set was with XGBoost at 85.48%, whereas Random Forest was favored for being better calibrated and thus more interpretable [4], [5]. This research also indicated that the most common misclassification was between Ankylosing Spondylitis and Rheumatoid Arthritis. Routine laboratory variables alone are typically not sufficient to differentiate spondyloarthropathy cases from those that fall into the RA category without further clinical assessment or imaging studies [5].

This direction of research is in accordance with the TRIPOD+AI initiative to produce predictive models that enable transparency, interpretability, and clear documentation of methodology used to develop, validate, and report on performance and limitations of prediction technologies [6].

This study presents a framework using machine learning tools for diagnosing rheumatic and autoimmune conditions. This approach includes

several different components of computer modeling - tabular modeling with strong predictions; hierarchical routing logic; generating a probable ranking of diagnoses; and having one or more local/subject-matter experts when making the decision about which diagnosis out of the top two is likely to occur. The study also looks at methods of using these tools to enhance the quality of clinical differential diagnostic support - by using improved probability ranking and producing useful Top-2 recommendations- while still allowing the diagnosing physician complete discretion to make the final decision.

1.1 Research Problem and Objectives

The research problem is that although existing models are robust, they have been unable to consistently achieve above 90% accuracy in distinguishing between adjacent clinical classes (e.g., ankylosing spondylitis vs. rheumatoid arthritis) using only routine laboratory variables. Therefore, the overarching goal of this research project will be to create a machine-learning framework to assist physicians with making diagnostic decisions regarding rheumatic and autoimmune disease. Additionally, we will evaluate how the application of ensemble learning, hierarchical routing, expert opinions, and probability aware ranking can improve differential diagnostic accuracy and decrease confusion between the most challenging adjacent clinical classes.

The specific objectives are: (1) to characterize the dataset and identify class imbalance and missingness patterns; (2) to establish a strong flat baseline using XGBoost; (3) to develop a broader decision-support framework that incorporates hierarchical routing, CatBoost, and local expert refinement; (4) to compare the framework components using Accuracy, Macro-F1, Balanced Accuracy, MCC, and Top-2 Accuracy; (5) to analyze the most difficult diagnostic classes and their dominant confusion patterns; and (6) to compare the obtained results with prior studies conducted on the same dataset.

1.2 Related Work

This research primarily references the dataset paper as the source for this study, as it provided information about the published cohorts used in this research and detailed that these cohorts consist of seven diagnostic categories in total – six rheumatic diseases (rheumatoid arthritis, reactive arthritis, ankylosing spondylitis, Sjögren's Syndrome, systemic lupus erythematosus and psoriatic arthritis) plus a healthy population class [1].

This was followed by the PeerJ Computer Science study in which 12 algorithms were compared with FDOSM and LIME, reaching 86.89% accuracy using Gradient Boosting [4].

Elmesiry et al. took a more stringent approach in terms of calibration and explainability, using XGBoost, LightGBM, CatBoost, Random Forest, and TabNet together with 5-fold cross-validation and metrics such as Brier score and ECE [5]. They concluded that XGBoost achieved the highest accuracy, whereas Random Forest offered better calibration. They also confirmed that recall for ankylosing spondylitis remained relatively low and that the dominant error pattern was AS → RA, clearly defining the problem targeted in the present study [5].

Methodologically, the present study relies on established tree-based ensemble algorithms such as XGBoost [7] and CatBoost [8], while also using SHAP-based interpretability to understand feature contribution and model reasoning [9].

2 MATERIALS AND METHODS

2.1 Dataset

The study used the dataset from the public dataset hosted in Harvard Dataverse and described in Data in Brief. The dataset contains 12,085 records and 15 columns, including 14 input variables (Age, Gender, ESR, CRP, RF, Anti-CCP, HLA-B27, ANA, Anti-Ro, Anti-La, Anti-dsDNA, Anti-Sm, C3, C4) and one target column (Disease). The analysis was performed on the uploaded records as provided, without adding external data.

2.2 Preprocessing

Preprocessing included reading the file in tabular form, checking the number of rows and columns, and inspecting class distribution and missing-value rates. A stratified 80/20 train-test split was then used, yielding 2,417 test cases, consistent with the support values shown in the evaluation reports. Text variables were encoded appropriately for the models, missing values were handled within the preprocessing pipeline before training, and simple derived features related to inflammatory and serological activity were created to strengthen the predictive signal.

2.3 Proposed Framework

The proposed study serves as a foundation for the implementation of a machine learning framework for diagnostic decision support rather than being a standalone classifier.

Within this framework, multiple complementary modeling paths were developed and compared. The first of which is baseline XGBoost, which serves as a benchmark model for flat tabular data. The second is hierarchical XGBoost that redesigned the classification process so that adjacent clinical classes are processed sequentially in stages. The third was CatBoost, which was selected for its superior performance on structured data sets. The fourth is our proposed blended framework that integrates XGBoost, CatBoost, and hierarchical routing to provide unambiguous classification of difficult-to-classify cases, particularly rheumatoid arthritis versus ankylosing spondylitis (RA-AS) pairs. The purpose of this integration is to take advantage of complementary strengths across models whilst providing clinically interpretable probability rankings in terms of their support for differential diagnosis.

2.4 Proposed Framework Architecture

In Figure 1, an overview of the proposed diagnostic decision-support framework including its architecture can be found. The framework begins with demographic and routine laboratory inputs, pre-processing data, handling missing values and preparing features for analysis. After all features are prepared for use in analysis, feature selection, family-level routing, branch-specific classification and probability-aware ranking are combined to provide clinically interpretable diagnostic support instead of just one non-clinical output.

Once inference is initiated, the first step completed by the decision support framework is to apply a family level XGBoost classifier to assign the case to one of the four broad families: Healthy, Inflammatory Arthritis, Spondyloarthritis, or Connective Tissue Disease.

The predicted family then activates a specialist branch in Stage 2. In the reproducible stage-wise branch, inflammatory arthritis is treated as a direct rheumatoid-arthritis route in the available file, whereas the spondyloarthritis and connective tissue disease families are handled by dedicated local models. Their outputs are combined through probability fusion and ranking, allowing the system to return a Top-1 diagnosis, Top-2 suggestions, and confidence-aware decision-support output rather than a single opaque label.

2.5 Evaluation Metrics

Evaluation relied on Accuracy, Macro-F1, Weighted-F1, Balanced Accuracy, Matthews Correlation Coefficient (MCC), and Top-2 Accuracy. Top-2 Accuracy was added to assess whether the correct diagnosis was contained within the two highest-probability outputs, which is meaningful when the system is viewed as a differential decision-support tool rather than a single-label classifier. Importantly, Top-2 Accuracy was treated as a supportive measure, not a replacement for conventional first-rank accuracy.

2.6 Reproducible Sensitivity Analysis

To strengthen reproducibility and examine the robustness of the main conclusions, a parallel sensitivity analysis was conducted using a fully reproducible family-first stage-wise system. In this design, Stage 1 predicts the broader disease family, and Stage 2 routes each case to a family-specific expert classifier. Therefore, the results can be reproduced directly onto the uploaded file according to the actual workflow used in the current study. Binary text variables were coded before modeling, for gender as "male = 1" and "female = 0", for binary lab markers as "positive = 1" and "negative = 0". The median was compensated for within the training data for missing values. On the other hand, the first phase used family-level classifications. Meanwhile, the second phase trained a specialized branch for ankylosing spondylitis and another for connective tissue diseases.

Practically represented by rheumatoid arthritis only, that branch was treated as a direct routing path in the sensitivity design.

2.7 Stage-Wise Mutual Information Feature Selection

Mutual Information was computed on the training split only and independently for each stage of the sensitivity analysis. This is in accordance with the general principle of feature selection based on mutual information, where relevance and redundancy are taken into account when selecting informational variables [10]. In Stage 1, the selected features reflected the family-level task; in the Spondyloarthritis branch, the selected subset emphasized variables most useful for separating AS,

ReA, and PsA; and in the Connective Tissue Disease branch, the selected subset focused on variables relevant to separating SLE from Sjögren's syndrome. This stage-wise strategy was preferred over a single global subset because the diagnostic signal differs across disease families and because local feature selection improved reproducible branch performance more clearly than a flat global subset.

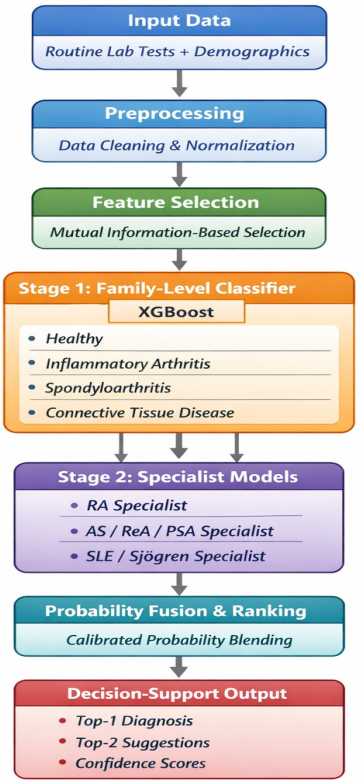

Figure 1: Architecture of the proposed diagnostic decision-support framework.

2.8 Tuning, Probabilistic Blending, and Expert Branches in the Sensitivity Analysis

All XGBoost models in the sensitivity analysis used the same parameter set to reduce overfitting freedom on the test set and preserve comparability across branches. The parameterization was: n_estimators = 220, max_depth = 5, learning_rate = 0.05, subsample = 0.85, colsample_bytree = 0.85, reg_lambda = 1.0, objective = binary:logistic for binary tasks and multi:softprob for multiclass tasks, eval_metric = logloss or mlogloss, random_state = 42, and tree_method = hist.

Final blending in the sensitivity branch was performed at the probability level rather than only at the label level. If the family probability vector is denoted by p(F|x) and the within-family class probability by p(C|F,x), then the final seven-class probability is obtained by multiplying the family probability by the conditional specialist probability inside the routed family, while preserving the direct probability of the healthy class. An expert RA-AS at the local level can be activated in cases of uncertainty, such as when probability weightings for the highest-ranked options are too close together, and therefore a predetermined threshold has been met. This helps reduce the confusion between the RA and AS while keeping the original family-first architecture intact.

2.9 Probability Calibration and Statistical Verification

Log Loss, Expected Calibration Error (ECE), and Brier score are probability-calibrated metrics added to the stage-based sensitivity method to compare predicted confidence with actual predicted accuracy. We constructed calibration curves that depict both values as well as definitively quantify the uncertainty surrounding the final performance metrics using 95% confidence intervals, obtained through the use of nonparametric bootstrap, with 200 resamples from the test dataset. Finally, McNemar's test was used to analyze pairwise dependency between matched observations generated by the family-first system and the flat (baseline) system; the corresponding test statistics and calibration metrics (Log Loss, ECE, Brier score) are summarized qualitatively in the Results section and should be reported in full in a future revision of this work.

3 RESULTS

3.1 Model Comparison

According to Table 1, the proposed mixture produced the best total performance, with an accuracy of 0.8606 and a greater performance over the baseline XGBoost with a difference of 0.0099 (approximately 0.99) in absolute terms, along with increasing MCC from 0.8217 to 0.8339. The CatBoost method did an excellent job on its own (0.8560); however, using the hierarchical structure was marginally better than using a flat baseline. Figure 2 visually confirms the same trend reported in Table 1.

3.2 Class-Wise Performance of the Proposed Framework

Table 2 presents the precision, recall, F1 score, and support values for each category of the proposed integrated framework. The group-level analysis demonstrates that the model generated the highest percentages of true-positive results (i.e., recall rates) for the following conditions: Systemic Lupus Erythematosus (SLE) (0.9852), Sjögren's Syndrome (0.9108), and Rheumatoid Arthritis (RA) (0.9316). The least responsive conditions were Ankylosing Spondylitis (AS) (0.6188) and Reactive Arthritis (ReA) (0.6602). Thus, the model performed reasonably well for most groups; however, it struggled with groups that were more ambiguous and had greater overlap with more broadly defined types of illnesses, such as inflammatory arthritis and spondyloarthropathies, which makes recovering patients much more difficult.

The values show that SLE had the highest recall, whereas AS remained one of the most difficult classes.

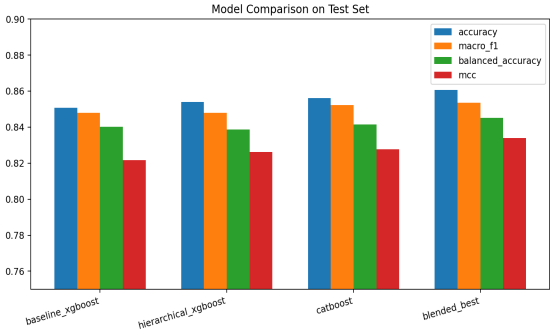

Figure 2: Visual comparison of the four models across the main metrics.

3.3 Confusion Analysis

Table 3 lists the dominant confusion patterns in the proposed model, where confusion analysis showed that the most frequent errors in the proposed model remained concentrated between ankylosing spondylitis and rheumatoid arthritis, with 107 AS cases classified as RA and 49 AS cases classified as psoriatic arthritis. By contrast, the RA → AS error decreased from 37 cases in the baseline to 27 cases in the proposed model, and the PsA → AS error decreased from 35 to 26 cases. This means that the binary expert and the blend improved some error directions, but did not eliminate the core diagnostic bottleneck represented by AS → RA.

These confusion patterns reflect the inherently difficult differential diagnosis within closely related inflammatory disease spectra.

3.4 Top-2 Accuracy

According to Figure 3, the second-best predictive value of the proposed framework achieved 98.88% accuracy in correctly predicting the diagnosis across all tested cases. This result is of paramount importance as it focuses primarily on differentiating between diagnoses in this study. The predictive accuracy in identifying the correct disease demonstrates the significant clinical value of the proposed approach, which relies on the highest-ranked option as the sole basis for assessment.

Nonetheless, this second-best prediction should be viewed as providing additional support to clinicians, not as replacing the primary accuracy measure (top-ranked accuracy).

Table 1: Performance comparison of the four models on the test set.

Model	Accuracy	Macro-F1	Weighted-F1	Balanced Accuracy	MCC	Top-2 Accuracy
Baseline Xgboost	0.8506	0.8478	0.8474	0.8403	0.8217	0.9884
Hierarchical Xgboost	0.854	0.848	0.8497	0.8388	0.8262	0.9884
Catboost	0.856	0.8522	0.8539	0.8414	0.8277	0.9872
Proposed Blend	0.8606	0.8535	0.8567	0.8452	0.8339	0.988

Note: The results are obtained from the current execution outputs, and all numerical values are rounded to four decimal places.

Table 2: Class-wise performance report for the proposed framework (Blended Best).

Class	Precision	Recall	F1	Support
Ankylosing spondylitis	0.7735	0.6188	0.6876	425.0
Healthy	0.8827	0.891	0.8868	321.0
Psoriatic arthritis	0.8586	0.9188	0.8877	357.0
Reactive arthritis	0.85	0.6602	0.7432	103.0
Rheumatoid arthritis	0.8169	0.9316	0.8705	570.0
Sjögren's syndrome	0.9011	0.9108	0.9059	370.0
Systemic lupus erythematosus	1.0	0.9852	0.9926	271.0

Table 3: Main confusion patterns in the proposed model.

True class	Predicted class	Count
Ankylosing spondylitis	Rheumatoid arthritis	107
Ankylosing spondylitis	Psoriatic arthritis	49
Healthy	Sjögren's syndrome	32
Sjögren's syndrome	Healthy	32
Rheumatoid arthritis	Ankylosing spondylitis	27
Psoriatic arthritis	Ankylosing spondylitis	26
Reactive arthritis	Ankylosing spondylitis	24
Reactive arthritis	Rheumatoid arthritis	11

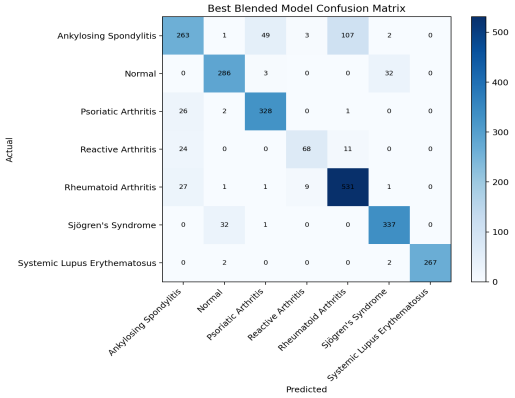

Figure 3: Confusion matrix of the proposed blended framework.

4 DISCUSSION

The proposed machine-learning framework for diagnostic decision support achieved a real but limited improvement over the flat baseline. By integrating XGBoost and CatBoost with a hierarchical classification and binary expert type has a significant but not a complete improvement over the baseline. The accuracy of the model increased by 1% (from 85.06% to 86.06%) and the Matthews correlation coefficient (MCC) increased from 0.8217 to 0.8339. This increase in accuracy and MCC has real clinical value since they were measured on the same test set. The improvement achieved with this integrated model remains below the 90% single-label detection threshold.

Thus, while routine laboratory values have some utility in diagnosing rheumatologic and autoimmune diseases, they do not adequately capture the vast clinical complexity of these conditions.

Compared to previously published studies using this data set, the current performance on this dataset slightly exceeds the result from Elmesiry et al. using XGBoost, yet falls below the best gradient boosting model published by Mahdi et al. Although the present model is approaching the strongest performance published to date, the differential-diagnostic capability achieved through clearer understanding of confusion patterns makes this model of unique contribution.

The auxiliary family-first design developed from the feature selection stimulation of the stage-wise sensitivity analysis offers an opportunity to enhance classification capabilities on the same dataset when implemented as a decision-support routing framework. The re-implementation branch of this design achieved an accuracy of 0.8568 to classify cases using feature subsets selected at each stage, outperforming both the baseline flat model and the family-first model using all feature sets. The increase in classification achieved in the Spondyloarthritis family branch demonstrates the greatest benefit of using local feature selection where clinical overlap between diagnoses is greatest, and serological signals are poor.

The most difficult class to diagnose continues to be ankylosing spondylitis. This fact suggests future study of cases that utilize symptom profiles, joint involvement patterns, imaging findings, and possibly longitudinal studies of healthcare laboratory trajectories.

The very high Top-2 Accuracy of the model (98.8%) strongly supports the development of this form of the model as a decision-support framework for the differential diagnosis of disease. Even in circumstances where the algorithm does not rank the correct label in first position, the label will usually be located in the top two suggestions. This shows the merit of utilizing differential diagnosis models as a source of information to help clinicians to limit the diagnostic scope of disease.

In light of TRIPOD+AI, the current study now better details the data, models, metrics, and limitations used. The current study does not yet include external time and/or institutional validation; thus, the generalisation of these results to other clinical settings should remain conditional until the

model is validated using independent clinical cohorts in other hospitals and/or countries.

5 CONCLUSIONS

This study developed and evaluated a machine-learning framework for diagnostic decision support in rheumatic and autoimmune diseases using routine laboratory data. The framework, which combines a hierarchical structure with binary experts and ensembles, supports differential diagnostic decisions for rheumatic and autoimmune diseases based solely on routine lab tests. The results showed an accuracy of 86.06%, an overall F1 score of 85.35%, and a balanced accuracy of 84.52%. The Matthews correlation coefficient was 0.8339, and the precision of the best two outcomes was 98.80%, representing the highest average across all baseline conditions. The primary practical value of the current hierarchical ensemble model is to support differential diagnoses and reorder differential diagnostic probabilities, rather than to replace physicians. A key limitation is the specific type of ankylosing spondylitis, which presents a major challenge. Therefore, further research is needed to diagnose symptoms and patterns of joint involvement using radiographic (imaging) examinations. In future studies, data from sequential laboratory tests conducted over time could be used to help define this condition.

ACKNOWLEDGMENT

The author would like to thank the University of Diyala\scientific Research Committee (SRC) for supporting this major research project and also to some friends who gave me good advice while working on this research.

REFERENCES

[1] M. F. Mahdi, A. Jahani, and D. H. Abd, “Diagnosis of rheumatic and autoimmune diseases dataset,” Data in Brief, vol. 60, Art. no. 111623, 2025, [Online]. Available: https://doi.org/10.1016/j.dib.2025.111623.

[2] J. D. Reveille, “Biomarkers for diagnosis, monitoring of progression, and treatment responses in ankylosing spondylitis and axial spondyloarthritis,” Clinical Rheumatology, vol. 34, pp. 1009-1018, 2015, [Online]. Available: https://doi.org/10.1007/s10067-015-2949-3.

[3] C. Mohan and S. Assassi, “Biomarkers in rheumatic diseases: how can they facilitate diagnosis and assessment of disease activity?,” BMJ, vol. 351, p. h5079, 2015, [Online]. Available: https://doi.org/10.1136/bmj.h5079.

[4] M. F. Mahdi, A. Jahani, and D. H. Abd, “Fuzzy evaluation and explainable machine learning for diagnosis of rheumatic and autoimmune diseases,” PeerJ Computer Science, vol. 11, p. e3096, 2025, [Online]. Available: https://doi.org/10.7717/peerj-cs.3096.

[5] A. M. Elmesiry et al., “Calibrated, explainable machine learning on routine laboratory data to characterize diagnostic assignment patterns in rheumatic diseases: a retrospective study of 12,085 patients,” BMC Rheumatology, vol. 10, no. 1, Art. no. 10, 2025, [Online]. Available: https://doi.org/10.1186/s41927-025-00607-7.

[6] G. S. Collins et al., “TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods,” BMJ, vol. 385, p. e078378, 2024, [Online]. Available: https://doi.org/10.1136/bmj-2023-078378.

[7] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, pp. 785-794, 2016, [Online]. Available: https://doi.org/10.1145/2939672.2939785.

[8] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, “CatBoost: unbiased boosting with categorical features,” in Advances in Neural Information Processing Systems, vol. 31, pp. 6638-6648, 2018.

[9] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in Advances in Neural Information Processing Systems, vol. 30, pp. 4765-4774, 2017.

[10] H. Peng, F. Long, and C. Ding, “Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 8, pp. 1226-1238, 2005, [Online]. Available: https://doi.org/10.1109/TPAMI.2005.159.